Screening for Myxobacterial Natural Products: New Structures, Biosynthesis, and Contributions to a Comprehensive Screening Workflow

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von
Thomas Hoffmann
Saarbrücken
2014

Tag des Kolloquiums:	18.07.2014
Dekan:	Prof. Dr. Volkhard Helms
Berichterstatter:	Prof. Dr. Rolf Müller
	Prof. Dr. Uli Kazmaier
	Prof. Dr. Christian Huber
Vorsitz:	Prof. Dr. Rolf Hartmann
Akad. Mitarbeiter:	Dr. Martin Frotscher

Diese Arbeit entstand unter der Anleitung von Prof. Dr. Rolf Müller in der Fachrichtung 8.2, Pharmazeutische Biotechnologie der Naturwissenschaftlich-Technischen Fakultät III der Universität des Saarlandes von Dezember 2009 bis Februar 2014.

Danksagung

An dieser Stelle möchte ich zunächst und ganz besonders meinen Eltern Gabi und Dieter, meinem Bruder Michael als auch Christina danken. Deren langjährige Unterstützung und Verständnis war ein wichtiger Bestandteil des Großprojekts „Studium und Promotion“.

Prof. Dr. Rolf Müller danke ich zutiefst für die herzliche Aufnahme in seine Arbeitsgruppe und die Chance die er mir damit eröffnet hat. Danke für das entgegengebrachte Vertrauen, die Übertragung von Verantwortung, die Unterstützung und die gute Zusammenarbeit.

Dr. Daniel Krug möchte ich besonders danken. Zum einen dafür, dass er stets auf mein Urteil vertraute und gleichzeitig jederzeit mit Rat und Tat zur Seite stand. Besonders hervorzuheben die unzähligen Stunden die er beim Verfassen wissenschaftlicher Texte geholfen hat. Danke für die gute Zusammenarbeit!

Dank auch an all die netten Kollegen die mich zu Beginn meiner Arbeit herzlich empfangen und in die, für mich neuen Themengebiete, eingeführt haben insbesondere Dominik Pistorius, Niña „Kookie“ Cortina und Ronald Garcia.

Und zu guter Letzt ein Hoch auf alle Kollegen und Projektpartner für die Zusammenarbeit und stete Bereitschaft über alles Mögliche zu reden, Fachliches zu diskutieren und bei Problemen zu helfen. Insbesondere all den Experten auf ihren jeweiligen Gebieten gebührt Dank für all die Hilfe, sei es im Hinblick auf Biosynthese, Bioassays, Fermentation, NMR, Mikrobiologe, etc.

Viele wurden zu Freunden.
In diesem Sinne, Danke für die gute Zeit!

Zusammenfassung

Mikrobielle Naturstoffproduzenten stehen aufgrund der zunehmenden Ausbildung von Antibiotikaresistenzen erneut im Fokus der Suche nach pharmazeutisch relevanten Wirkstoffen. In diesem Kontext befasst sich die vorliegende Arbeit sowohl mit der Charakterisierung neuer myxobakterieller Naturstoffe und deren Biosynthese als auch mit der Entwicklung einer analytischen Methode zur Analyse komplexer biologischer Extrakte.

Im Rahmen dieser Arbeit wurden die Naturstoffe Pellasoren und Microsclerodermin aus dem genomsequenzierten Stamm *Sorangium cellulosum* So ce38 isoliert. Neben der Strukturaufklärung des Pellasoren konnte ein zugehöriges Biosynthese-Gencluster identifiziert und dadurch die Biosynthese aufgeklärt werden. Zudem wurden in einem umfangreichen Screening-Projekt mehrere terrestrische Myxobakterien als Produzenten des marinen Naturstoffes Microsclerodermin identifiziert. Dies ist einer der wenigen bekannten Fälle der Auffindung identischer Sekundärmetabolite aus marinen und terrestrischen Quellen. Die Isolierung zweier neuer Derivate führte schließlich zur Charakterisierung der zugehörigen Biosynthesegene in zwei myxobakteriellen Stämmen. Dabei konnten die strukturellen Unterschiede der gefundenen Derivate durch die Unterschiede im Biosynthesecluster erklärt werden.

Des Weiteren wurde eine Methode zur Steigerung des Informationsgehalts von LC-MS/MS Daten entwickelt. Durch statistische Auswertung von LC-MS Datensätzen können Signale, welche aus dem Metabolismus des Bakteriums hervorgehen, identifiziert und daraufhin unabhängig von der jeweiligen Signalintensität fragmentiert werden. Die Anwendung dieser Methode führte zur Auffindung und Aufklärung der Struktur der Lipothiazole, einer neuen Substanzklasse aus *S. cellulosum* So ceGT47.

Abstract

The rapid emergence of antibiotic resistance enforces the search for new pharmaceutically relevant compounds. Here, especially natural products from microbial sources are becoming important once again. In this regard, the present work deals with the characterization of new myxobacterial natural products and their biosynthesis alongside with the development of an improved method for the chemical analysis of complex biological samples.

This thesis covers the isolation of two natural products, pellasoren and microsclerodermin, both derived from the crude extracts of the genome-sequenced strain *Sorangium cellulosum* So ce38. Full structure elucidation of pellasoren led to the identification of its biosynthetic gene cluster and the underlying biosynthesis. Furthermore, as a result of a comprehensive screening project, several terrestrial myxobacteria were unambiguously identified as producers of the marine natural product microsclerodermin. This finding resembles one of the few known cases of an identical metabolite derived from terrestrial and marine habitats. The isolation and subsequent characterization of two new derivatives from different strains resulted in the identification of two biosynthetic gene clusters. As a consequence, structural differences of the derivatives could be rationalized based on slight variations within both clusters.

In addition, a method to optimize the informational content of LC-MS/MS data was developed. Statistical evaluation of LC-MS data facilitated the identification of signals linked to bacterial metabolism, which were subjected to MS/MS analysis irrespective of their intensity. This methodology was successfully applied to identify the novel lipothiazole compound class in *S. cellulosum* So ceGT47.

Vorveröffentlichungen zur Dissertation

Teile dieser Arbeit wurden vorab mit Genehmigung der Naturwissenschaftlich-Technischen Fakultät III, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen veröffentlicht. Es handelt sich dabei jeweils um Erstautorenschaft.

C. Jahns[+], **T. Hoffmann**[+], S. Müller, K. Gerth, P. Washausen, G. Höfle, H. Reichenbach, M. Kalesse, and R. Müller; Pellasoren: structure elucidation, biosynthesis, and total synthesis of a cytotoxic secondary metabolite from Sorangium cellulosum; *Angewandte Chemie International Edition*, **51** (21), 5239–5243, 2012.

DOI: 10.1002/anie.201200327

C. Jahns[+], **T. Hoffmann**[+], S. Müller, K. Gerth, P. Washausen, G. Höfle, H. Reichenbach, M. Kalesse, and R. Müller; Pellasoren: Struktur, Biosynthese und Totalsynthese eines zytotoxischen Sekundärmetaboliten aus Sorangium cellulosum; *Angewandte Chemie*, **124** (21), 5330–5334, 2012.

DOI: 10.1002/ange.201200327

T. Hoffmann, S. Müller, S. Nadmid, R. Garcia, and R. Müller;

Microsclerodermins from Terrestrial Myxobacteria: An Intriguing Biosynthesis Likely Connected to a Sponge Symbiont; *Journal of the American Chemical Society*, 2013.

DOI: 10.1021/ja4054509

Table of Contents

1 Introduction

1.1 The Role of Natural Product Research

Throughout history humans have been suffering from numerous infectious diseases. Especially infant and child mortality is strongly linked to the occurrence of pathogens as causative agents of infections as they exploit the rather weak immune response during the first months after birth. Even among adults infectious diseases remain to be a major threat to health. There are approximately 200 infectious diseases caused by bacteria, viruses, fungi, protozoa, parasites and prions whereas the most part is related to bacteria and viruses. It is noteworthy that only a small subset of all bacteria is harmful to the human organism albeit those can cause severe health problems or even death to the infected person. With the rise of antibiotic application in the 19th century, humans managed to fight against such infectious diseases for the first time in a broadly applicable manner. There has been always certain knowledge of how to heal infections; a knowledge usually based on elderly wisdom or witchcraft based solely on trial and error and the cultural heritage evolved thereof.[1] However, the underlying principles for the observed healing effect were not understood, although used successfully. This changed with the understanding that microorganisms are causative for infections as first postulated by Jakob Henle and Edwin Kleb and finally proven by Robert Koch and Friedrich Löffler.[2,3]

The use of antibiotics gained momentum in the early 20th century when scientists realized that certain microbes are able to inhibit the growth of other microorganisms.[1] Finally, the steadily increasing knowledge in chemistry resulted in the identification of single compounds being responsible for the observed effects, e.g. the widely renowned penicillin which was the first isolated natural product showing an antibiotic effect. It represents the archetype of a natural product bearing antimicrobial activity. Nowadays, natural products are by definition chemical compounds produced by a living organism that have biological or pharmacological activity. As many infectious diseases are caused by bacterial or fungal microorganisms that may be per se harmful or exploit at least opportunistic pathogenicity, a suitable way to tackle such infections is to kill the pathogenic bacterium (bactericidal activity) or stop its growth (bacteriostatic activity). A reason for the production of compounds exhibiting antibiotic properties could be a defense mechanism against other microorganisms, presumably as a consequence of the dense communities microorganisms live in and the resulting cope for nutrients and living space. According to this circumstance, it is not surprising that many microbes are equipped with the genetic capabilities to produce molecules that can affect the life of other microbes. Such compounds are not required for the basic metabolic processes of life; hence, they are named secondary metabolites to emphasize their role in

metabolism. It is noteworthy that secondary metabolites do not exhibit antibacterial activity per definition. In particular, for many secondary metabolites it remains unknown for which reason they are produced by the respective microbe.[4]

In the end, the number of natural products is literally uncountable since the biosynthesis of natural products evolves together with the organisms, thereby creating a plethora of different chemical structures. Some of those structures were the basis for the antibiotic era of mankind which started around 100 years ago and led to a previously unseen gain in life quality and extension of human life time. The important role of natural products becomes obvious in light of the drugs that are derived from natural products (Figure 1).[5]

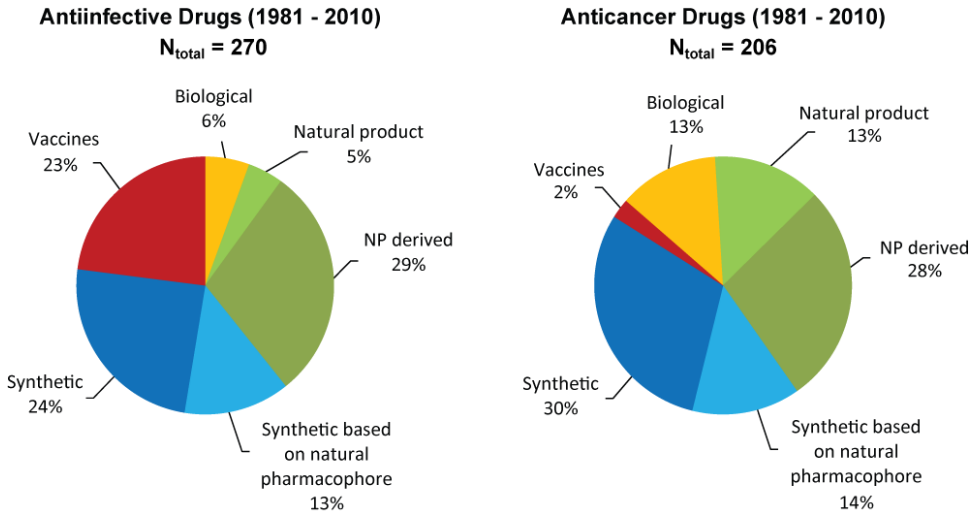


Figure 1: Approved anti-infective and anticancer drugs categorized according to their origin. Graphic adapted from Newman and Cragg.[5]

However, the current trend of increasing resistance against available antibiotics amongst pathogenic bacteria could easily end in a new pre-antibiotic era.[6,7] Some pathogenic bacteria are now even pan-resistant, i.e. resistant against all available antibiotics. Among all pathogenic bacteria a group of seven is regarded as highly problematic. These so-called ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumonia*, *Acinetobacter baumanii*, *Pseudomonas aeruginosa*, and *Enterobacter* species) do frequently "escape" the available antibiotics and eventually lead to death of the patients.[8] The situation demands to use again old drugs that were not longer on the market due to their toxic side effects. Some of these drugs such as colistin are the last resort against gram-negative multi-resistant bacteria.[9] The only way to overcome this scenario lies in new antibiotics reaching the market. Novel chemical scaffolds are necessary to identify new mode of actions that circumvent current resistance mechanisms. Since only a low percentage of new natural products make it to the stage of clinical trials and even less finally are approved as drugs it is obvious that there is a huge demand for new compounds to furnish the pharmaceutical pipelines. In conclusion, the discovery and development of

new natural products are important for human health. However, the "low-hanging fruits" among antimicrobial natural products have already been harvested. It is therefore important to carefully choose promising sources for further investigation. Such sources are usually found in biological niches or in habitats that are rather underexploited. In particular, this understanding led to an increased interest in the rich marine habitats and in uncommon types of bacteria and fungi, following the current notion that chemical diversity comes along with phylogenetic diversity. The myxobacteria are one of these underexploited types of bacteria that have been steadily yielding new natural products with promising modes of action.[10]

1.2 Myxobacteria – a Special Group of δ-Proteobacteria

Within the class of δ-proteobacteria the order *Myxococcales* represents an uncommon type of gram-negative bacteria. These myxobacteria show an intriguing "social behavior" and a complex life cycle. Although myxobacteria are single cell organisms, they are able to form multi-cellular fruiting bodies that contain spores to endure times of limited nutrition supply or unfavorable environmental conditions. Fruiting body morphology and spore shape are useful as classifiers to characterize the different myxobacterial families.[11] Myxobacteria are also able to move on surfaces by swarming and creeping in order to find nutrient resources, which is certainly a remarkable type of behavior for proteobacteria. Preferred nutrients are macromolecules that are lysed by suitable exo-enzymes, but some strains are even able to prey on live microorganisms such as yeast or *E. coli*.[12]

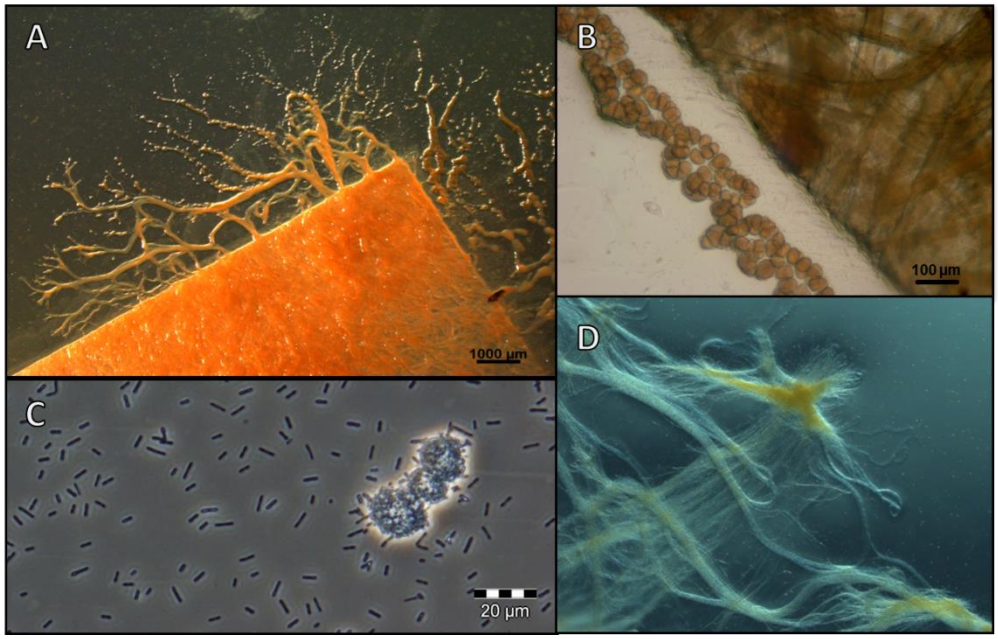


Figure 2: ***A*** *Stereophotomicrograph of Sorangium cellulosum radiating swarm from edges of filter paper.* ***B*** *Brightfield photomicrograph of* Sorangium cellulosum *fruiting bodies at the edge of filter paper.* ***C*** *Rod-shaped, blunt end vegetative cells of Sorangium cellulosum So ce38 in liquid medium.* ***D*** *Stereophotomicrograph of Jahnella sp. MSr9139 swarming on agar.(Pictures by Ronald Garcia)*

The fascinating characteristics of myxobacteria are linked to the complex interplay of intercellular signaling, cell differentiation and morphological changes.[13,14] Some of these processes, such as the ability to move by gliding require the interaction of numerous proteins.[13] The fact that myxobacteria seem to have the largest genomes amongst bacteria sounds reasonable in light of the aforementioned capabilities.[15]

Most myxobacteria isolated to date are of terrestrial origin although a few marine strains have been characterized.[11,16,17] Marine habitats are currently underexploited in terms of screening for myxobacteria, thus the low numbers of known marine strains could be attributed to insufficient sampling efforts. However, phylogenetic analysis of marine sponges based on 16S rRNA analysis suggests that myxobacteria may be present in sponges and marine sediments.[18,19] Although this evidence is still awaiting ultimate confirmation by isolation and cultivation of sponge-derived myxobacteria, it points towards a huge potential resource of new marine myxobacteria, providing much-desired phylogenetic diversity for future screening campaigns. Past and ongoing screening efforts have revealed myxobacteria as a valuable source for new natural product classes that frequently show promising modes of action.[10] This biotechnological potential is partly owing to the exceptional size of myxobacterial genomes.[20] Comprising frequently more than 10 Mio. base pairs, the genomes of the genus *Sorangium cellulosum* seem to be the largest bacterial genomes with sizes of 13.0 Mbp for strain So ce56 and 14.8 Mbp for strain So0157-2 .[15,21] In addition to their size, myxobacterial genomes are uncommon as they feature a high GC content of around 70 %.[15] Currently, several myxobacteria are sequenced and primary results indicate similar genome sizes and GC contents.

Interestingly, the majority of the biologically active secondary metabolites act against fungi or bacteria, which are common competitors of myxobacteria in their natural habitat.[10] For example, a cellulose-degrading *S. cellulosum* strain may combat cellulose-degrading fungi by producing antifungal compounds whereas proteolytic myxobacteria benefit from antibacterial compounds that can act on other proteolytic competitors or prey.[22,23] Other compounds have cytotoxic, anti-viral, or anti-inflammatory activity, which do not per se confer an evolutionary advantage to the producing strain and may be a serendipitous effect of the compound. This can be attributed to the limited number of protein folds that are known. At the time of writing this thesis, the SCOP database listed only 1360 different fold types for proteins.[24] In view of this circumstance, it seems quite obvious that some compounds can interact simply by chance with a protein. There are also compounds showing no activity against other organisms when tested with standard bioactivity assays. Those compounds may have other functions, e.g. such as myxochelin which is a siderophore and helps iron acquisition by chelating iron, or they are involved in signaling pathways, or they are simply not active in the limited collection of assays that are available.[25] The metabolic effort to produce secondary metabolites is significant and it is considered unlikely that they would be produced without being directly or indirectly beneficial for the producer.[4,26]

This should be considered when a new compound is characterized that lacks activity on first sight. Not all secondary metabolites are per se bioactive under pharmaceutically relevant conditions and for many of them the type of bioactivity needs to be elucidated first.[26] In light of this, it is noteworthy that the combination of new assays and a tight collaboration between research groups is beneficial for finding more promising activities.

Research of the last decades has revealed that myxobacterial secondary metabolites often have uncommon or previously undescribed modes of actions, making them particularly interesting for pharmaceutical research.[10] Among the myxobacteria screened so far, the most proficient sources for natural products are strains of *Sorangium cellulosum*, *Myxococcus*, and *Chondromyces*.[22] Strains of *S. cellulosum* are classified in the family *Polyangiaceae* of the suborder *Sorangiineae*. These cellulose degrading strains are usually terrestrial isolates derived from decaying plant material. Some of these strains are moderately thermophilic and grow at 42-44 °C while the majority is culturable at 30 °C.[11,27] Many potent natural products were isolated from strains of the *Sorangiineae* suborder, e.g. ambruticin[28], epothilon[29], ripostatin[30], sorangicin[31], soraphen[32], and thuggacin[33] (Figure 3). The majority of myxobacterial secondary metabolites reported to date are derived from huge protein complexes of the polyketide synthase (PKS) type, the nonribosomal peptide synthetase (NRPS) type, or hybrids thereof (see Chapter 1.3). In addition, current research reveals an increasing number of ribosomally produced natural products (RiPPs).[34] However, their overall number is still moderate with one possible reason being that RiPP biosynthetic gene clusters circumvent a straightforward detection by genome mining approaches.

Over the last decades, research with myxobacteria yielded around 900 compounds of 140 compound classes isolated and characterized at the Helmholtz Center for Infection Research, Braunschweig, Germany and at the Helmholtz Institute for Pharmaceutical Research, Saarbrücken, Germany (as of 01/2014). The overall number of myxobacterial natural products is quite remarkable in light of the rather small number of strains screened (< 10,000), especially when compared to the screening efficiency in actinomycetes research.[35] However, myxobacterial natural products still account for only a low percentage of all published microbial natural products. Most of these are derived from actinomycetes, fungi, and bacilli strains;[22] a fact owing to the long-standing research in these fields in combination with extensive screening programs conducted by industrial and academic research groups.

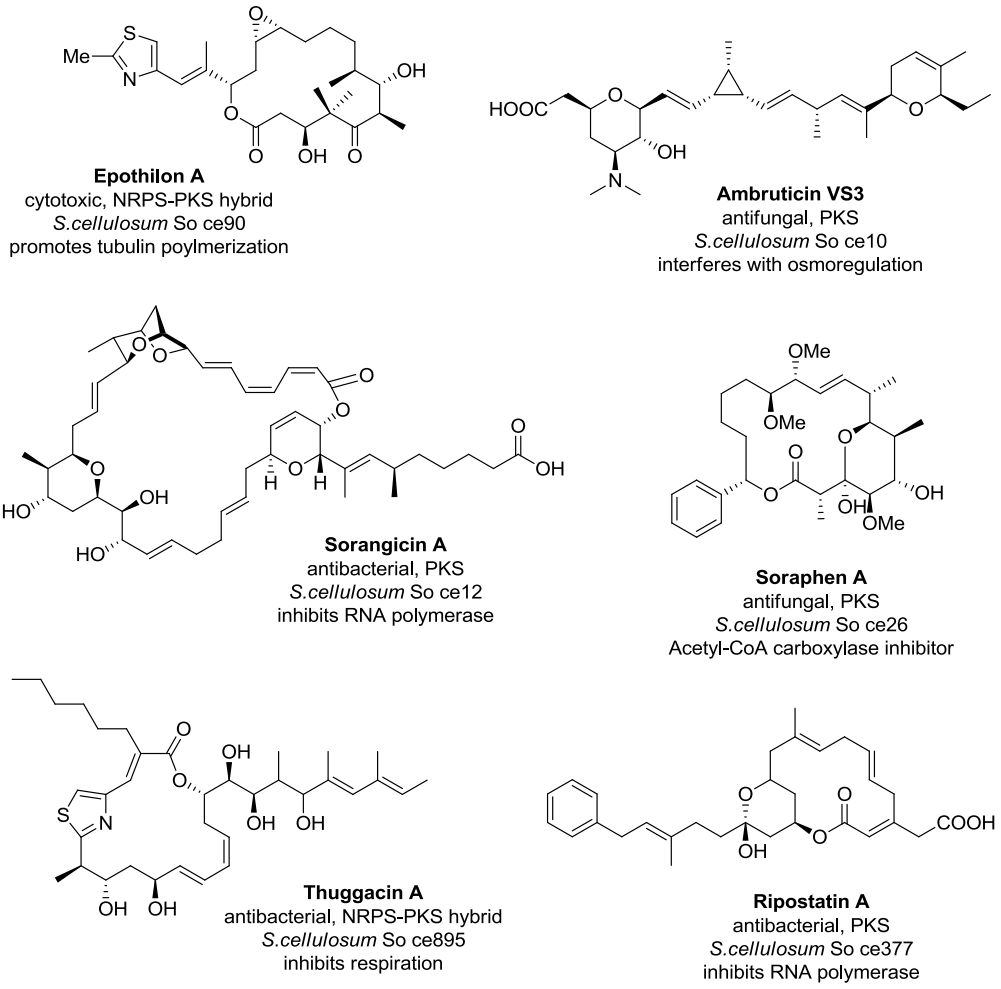


Figure 3: Selection of bioactive secondary metabolites isolated from Sorangium cellulosum strains.

1.2.1 Discovery of New Secondary Metabolites from Myxobacteria

The current notion of how to exploit myxobacterial secondary metabolism comprises several fundamental ideas, i.e. (1) isolation of new myxobacterial strains, (2) using state-of-the-art instrumental analytics in combination with (3) a versatile screening for bioactivity against a panel of representative indicator strains and (4) applying the full range of advanced genome-related techniques such as genome mining, directed mutagenesis, and heterologous expression of biosynthetic pathways in alternative hosts.

Ongoing efforts to sequence myxobacterial genomes continue to reveal a huge potential for secondary metabolite production.[36] So far, numerous loci within each sequenced genome harbor genes that belong to secondary metabolite biosynthetic gene clusters. The number of such clusters varies from around 10 to > 20 clusters per strain. On the contrary, only a subset of the corresponding compound classes is usually found in a crude extract. This mismatch may be attributed in part to the cultivation conditions, which hardly resemble a natural environment. Some compounds may be not produced at all, e.g. until an unknown signal initiates transcription of the responsible genes. Others may be produced in very low amounts and missed in typical screening approaches. However, at least one analysis of the proteome of *M. xanthus* DK1622 revealed that proteins for 17 out of 18 secondary metabolite gene

clusters were expressed.[37] In terms of cultivation conditions it was recently shown, that a strict control of the CO_2 and O_2 partial pressures during fermentation can influence productivity of different compounds.[38] Such effects are not considered if shaking flask cultures are prepared. After all, the reason why so many potential secondary metabolites are still not found remains elusive and achievements to induce production by different cultivation conditions were merely serendipitous. An interesting approach towards altering the secondary metabolome is co-cultivation of bacterial strains and fungi in order to induce unforeseen effects.[39,40] With respect to myxobacteria the effects of co-cultivation are to date still underexploited.

Table 1: Overview of some myxobacterial genome sizes, secondary metabolite clusters found in silico, and the number of clusters that were already linked to a secondary metabolite.

Strain	Genome [Mbp]	Secondary metabolite cluster overall	Secondary metabolite cluster assigned	Ref.
Myxococcus xanthus DK1622	9.2 Mbp	18	11	[41]
Sorangium cellulosum So ce56	13.0 Mbp	13	4	[15]
Stigmatella aurantiaca DW4/3-1	10.3 Mbp	14	5	[42]
Sorangium cellulosum So ce38 [a]	14.5 Mbp	8	2	this work
Sorangium cellulosum So ceGT47 [a]	10.9 Mbp	9	3	this work

[a] unpublished results

For *M. xanthus* DK1622 eleven out of 18 compound classes were identified under culture conditions so far. For *S. cellulosum* So ce56 compound classes were assigned to only 4 out of 8 gene clusters identified by genome sequencing. The projects of this thesis deal with two *S. cellulosum* strains, So ce38 and So ceGT47, which were sequenced at the starting point of this work. Table 1 lists assigned and unassigned secondary metabolite clusters of different genome-sequenced myxobacteria.

Several approaches address the apparent discrepancy between observed metabolite profiles and the genetic capacity for the production of secondary metabolites by introducing new methods for identifying secondary metabolites and new ways of using the genomic information. Among the genomics-based methods, perhaps the most straightforward is to create knock-out mutants by targeted disruption of a biosynthetic gene. This approach has been used with both site-directed and random transposon mutagenesis techniques for genetic manipulation of the organism under investigation.[43–45] Upon successful knock-out of a biosynthetic gene cluster the corresponding natural product cannot be produced anymore. Careful comparison of a wildtype strain's LC-MS chromatogram with the knock-out strain's chromatogram can help to identify the signals representing the compound produced by the inactivated gene cluster. However, this method requires first of all that the secondary metabolite is produced in the wildtype strain and demands at the same time for the development of a genetic manipulation method, a challenging undertaking with some myxobacteria. Moreover, the knock-out of a

biosynthetic gene cluster that is not expressed in the wild-type strain will obviously not result in an informative change between the metabolite profiles of wildtype and knock-out mutant. Even if all prerequisites are fulfilled it is tedious and error-prone to identify differential signals in complex LC-MS datasets by manual comparison. Peak intensities of metabolites and matrix components can be subject to strong variation among replicate cultivations, and unforeseeable changes may occur when cultivating the knock-out strain in the presence of an antibiotic to maintain the selection marker introduced for gene inactivation. Less abundant compound peaks are easily missed under such circumstances. This problem can be overcome by employing software-aided data evaluation e.g. by making use of principal component analysis (PCA) to highlight differences between two sample sets (see chapter 1.4).[46]

A gene cluster can be linked to a secondary metabolite by approaching the analysis from the opposite way, i.e. starting from a characterized secondary metabolite. Retrobiosynthetic considerations help to create a draft biosynthesis gene cluster that, in theory, could hold responsible for synthesizing the core structure of the metabolite. In a next step, all biosynthetic gene clusters of a producer strain's genome are compared to this draft in order to identify the one that fits best. This type of analysis was done for all compounds of this thesis as well as for the identification of other myxobacterial compounds, e.g. gephyronic acid and althiomycin.[47,48] This metabolite-to-genes approach can also be applied in a very early stage research when LC-MS/MS data of crude extracts is acquired. Here, indicative fragmentation patterns can highlight amino acid building blocks, which are then in turn mapped to the adenylation domains found within the genome, a method which was recently named peptidogenomics.[49]

Another method to discover and study new secondary metabolites uses heterologous expression of secondary metabolite gene cluster in suitable hosts, e.g. *Pseudomonas putida* or *Myxococcus xanthus*.[50,51] However, the size of NRPS or PKS gene clusters demands new techniques for working with large DNA fragments.[52,53] In addition, the number of suitable host organisms is limited owing to the high GC content of myxobacteria and the fact that some host organisms may lack metabolic pathways for the biosynthesis of uncommon precursors and for the post-translational activation of multi-domain biosynthetic enzymes.[36] Nevertheless, heterologous expression provides an opportunity to evade the weak point of several other approaches, with regard to low productivity of unknown metabolites as a consequence of inappropriate cultivation conditions, unforeseen regulation, and genetic inaccessibility.[36] However, similarly to the knockout- and comparative profiling approach, heterologous expression also requires the sensitive detection of potentially subtle differences between LC-MS samples.

Finally, when the assignment of a new metabolite to a biosynthetic gene cluster could be accomplished, the next challenge is to improve yield in order to enable compound isolation and full characterization, since newly discovered compounds may be initially produced in very low yields only. A good option is to isolate alternative strains that are better producers of a target compound. While the diversity of metabolite profiles to date has not been comprehensively revealed at large scale in terms of

spanning a large group of organisms, the analysis of increasing numbers of strains nevertheless seems to indicate for myxobacteria that profiles from related genera exhibit a certain overlap. However, production rates for specific compounds may differ by several orders of magnitude between the producing strains.[54] Obviously, a newly identified metabolite is easier to isolate if a good producer is available. Effectively using alternative producers from a wide choice of sources does however depend on availability of a large-scale analytical framework, capable of generating and handling quantitative information on production profiles and ideally covering an entire strain collection.

With respect to these requirements, a screening project for myxobacteria was initiated. The large collection of myxobacterial strains and compounds at the Helmholtz Center for Infection Research, Braunschweig, Germany and at the Helmholtz Institute for Pharmaceutical Research, Saarbrücken, Germany was the perfect starting point to establish a comprehensive screening platform. All known compounds as well as approximately 2500 crude extracts of carefully selected myxobacterial strains were measured with standardized conditions using this analytical platform (LC-MS) and furthermore checked for bioactivities. All information is gathered in our proprietary data base, the Mxbase, which allows a convenient access to all information that is available to any strain or compound. Such a framework, which aims for comprehensiveness, facilitates dereplication of known compounds and identification of new compounds as well as the above mentioned search for alternative producers.

1.3 The Rationale behind NRPS- and PKS-based Natural Product Biosynthesis

Many natural products are derived from nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS). Both biosynthetic systems use rather large, multi-modular protein complexes, which have been heavily studied for the last 20 years. In contrast, another source of natural products based on the ribosomal formation of short genome-encoded precursor peptides and extensive post-translational modification has been increasingly investigated only since a couple of years. These ribosomally produced natural products (RiPPs) play a steadily increasing role in natural product research.[34] However, RiPP biosynthetic genes are not as easy identified by genome mining approaches as NRPS or PKS clusters; hence, it is currently still challenging to predict the genetic capabilities of a given strain in this respect, apart from the previously reported RiP pathways. In the following paragraphs the discussion of biosynthetic logic for the production of myxobacterial natural products will focus on NRPS and PKS biosynthesis, since RiPPs are not within the scope of this work..

1.3.1 Polyketide Synthases – PKS

The name polyketide reflects the structural basis common to this type of natural product. A polyketide chain can be described as a regular polymer consisting of C2 units with a ketone at every second carbon. This holds true for the most basic type of polyketides s derived solely from a sequence of elongation steps without further reductive processing. There is essential similarity between polyketide chains and fatty acid chains and indeed, the proteins forming polyketide synthases (PKSs) are homologues of the fatty acid synthase (FAS) machinery. Progress in FAS research was of great benefit to understand the principles of PKSs,[55–57] although a PKS was not identified or characterized on genome or protein level for a long time.[58] Finally, around the beginning of the 1990s – almost 40 years after the first biosynthetic proposal by Birch *et al.* – it was evident that all polyketides are produced by the same biosynthetic principles as found in fatty acid synthesis.[59–61] Today it is proven that PKSs are huge megasynthases (up to > 1 MDa in size) of a homodimeric, twisted structure that share catalytic domains within the dimer.[62,63] Both FAS and PKS systems use a pool of simple precursors such as an acetyl-ester of coenzyme A (acetyl-CoA) and a malonate-ester of coenzyme A (malonyl-CoA) to build up carbon chains by C-C bond formation. Several distinct enzymes are necessary to perform all the steps of precursor acquisition, chain initiation, elongation, and further reductive processing. [58,64] The central biosynthetic principle is a Claisen condensation when the enzyme-bound malonate decarboxylates whereupon the electron pair undergoes a nucleophilic attack to the enzyme-bound acetyl moiety (Figure 4 A).

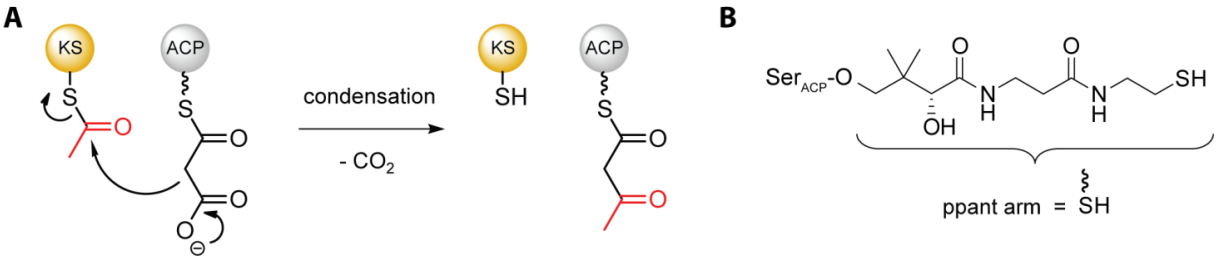


*Figure 4: **A** Central condensation step as observed in fatty acid and polyketide biosynthesis. **B** Structure of the phosphopantetheine arm attached to a serine of the acyl carrier protein. KS, ketosynthase; ACP, acyl carrier protein.*

The acetyl-CoA is used as a starter and delivered as an acetyl group onto the ketosynthase (KS). More precisely, an acyl transferase (AT) gathers the acetyl by transesterification of the coenzyme A ester with the hydroxyl group of a serine as depicted in Figure 5. This serine-bound acyl group is now transferred to the cysteine of the KS domain. Following the same principles, the malonyl group is attached to a thiol of the acyl carrier protein (ACP) via an acyl transferase. This thioester is formed with the terminal thiol of a phosphopantetheine moiety that is posttranslationally attached to a serine of the acyl carrier protein (Figure 4B). The phosphopantetheine (ppant) has the function of a flexible arm, which carries the growing chain to the different catalytic domains that act on the biosynthetic intermediate.[65] Hence, the intermediates can reach close proximity to the active centers of the assembly line, which plays an

important role for the efficiency of the whole process. After condensation of both units, the ACP-bound intermediate is further processed starting by reduction of the β-keto ester to form a β-hydroxyl ester, catalyzed by a keto reductase domain (KR). Following dehydration by a dehydratase domain (DH) and reduction of the double bond by an enoyl reductase domain (ER) the chain is ready for another round of elongation (Figure 5). For a fatty acid synthase, these subsequent steps are repeated until a certain chain length is reached upon which the product is cleaved off the ACP by a thioesterase (TE).

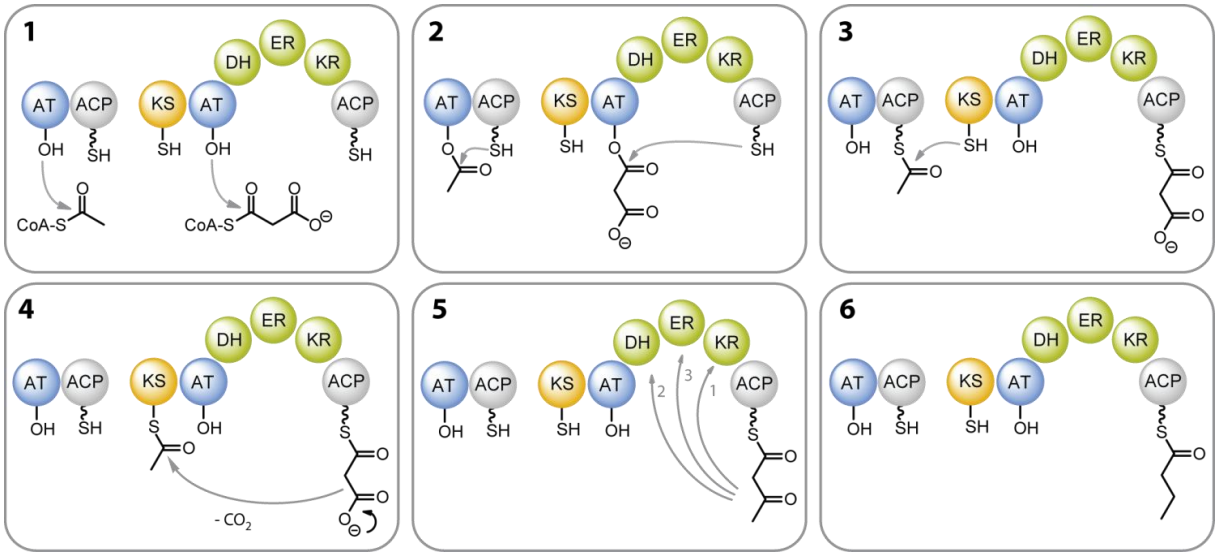


*Figure 5: Schematic reaction steps for one elongation step as found in FAS and PKS type I. The additional AT-ACP module is not found in FAS but usually present as loading module in PKS type I. (**1**) Loading of an activated precursor; (**2**) Transfer to a acyl carrier protein (ACP) by transesterification with ppant; (**3**) The starter unit is forwarded to a ketosynthase (KS); (**4**) Condensation reaction mediated by the KS domain; (**5**) Ketoreduction, dehydration, and enoyl reduction of the ACP-bound product; (**6**) Fully reduced chain extended by a C2 unit.*

While these basic enzymatic steps are the same for FAS and PKS type I, the way how enzymatic domains act together is different. A fatty acid synthase does always contain the same domains (KS, ACP, AT, KR, DH, ER, TE) performing iteratively as described above, whereas the biosynthetic logic of polyketide synthases is more flexible. PKS "assembly lines" are usually built up by single modules that often perform only one specific round of chain elongation. A PKS module is thereby defined as a set of enzymatic domains needed to extend the chain by a C2 unit. Hence, the minimal module is given by the domain set ACP-KS-AT, which will afford a β-keto ester. In addition to these essential domains a PKS module may contain further processing domains like KR, DH, and ER domains in order to further reduce the enzyme-bound β-keto moiety to a β-hydroxy ester, an α-β-unsaturated ester, or finally a saturated ester.[66] Moreover, there are various starter and extender units possible as a consequence of AT domain substrate specificity. The most common extender unit substrates are malonyl- (M-CoA) and methylmalonyl-CoA (mM-CoA). But also extender units based on ethylmalonyl-CoA, hydroxymalonyl, methoxymalonyl, and aminomalonyl have been reported.[67–71] The latter building blocks are not available from primary metabolic pathways, and thus are usually produced by specialized enzymes. To increase the repertoire,

several starter units such as propionate, isovalerate, phenylacetate, benzoate, cinnamic acid, amino acids, and many more be incorporated as reviewed by Moore and Hertweck.[72] The different specificities for starter unit or extender unit are governed by conserved motifs found in AT domains.[73] Directed modification of these motifs by genetic engineering in principle opens up the possibility to alter AT domain substrate specificity.[74,75]

Polyketide synthases are historically classified in three major groups: type I, II and III.[76] The initial constraints of the three groups did not last for long as more and more exceptions were found. As a result of the ongoing research, there are many "PKS type I" known that act iteratively, which was initially only attributed to type II PKSs (Chapter 1.3.4). PKSs of trans-AT type are another example that does not fit into the classification scheme. Hence, it is now deemed rather inappropriate to force PKS systems into these groups.[77] Indeed, PKSs are very versatile megasynthases made from a pool of different functions that allow numerous combinations and thereby bring about an impressive diversity of assembly line architectures. The flexibility of multimodular PKS architecture in turn gives rise to a plethora of polyketides. Moreover, PKS assembly lines may be extended beyond the basic set of enzymatic functions by additional enzymes, such as methyl transferases acting on carbon, nitrogen, and oxygen or aminotransferases just to name two classes. Those additional enzymes may be located inside modules, or alternatively be distinct enzymes that are encoded by genes downstream or upstream in respect to the PKS cluster.

1.3.2 Nonribosomal Peptide Synthetases – NRPS

An alternative way of enzymatic peptide synthesis apart from the ribosomal system was not known until the work of Lipmann *et al.*in 1971.[78] Today, nonribosomal peptide synthetases (NRPS) are a well-studied system that is found in many bacteria and fungi. Like PKSs, NRPSs are megasynthases producing a wide variety of peptide-derived secondary metabolites by a sequence of chain elongation and modification events. Each elongation step is again performed by a single module, which minimally a condensation domain (C), an adenylation domain (A), and a peptidyl carrier domain (PCP). From a general point of view, the minimal NRPS module resembles the domain triad AT-KS-ACP of PKS systems (Chapter 1.3.1), but the distinct biochemical steps catalyzed by NRPS domains are different. The biosynthetic logic of a NRPS is based on an activation of monomers by conversion of the free amino acid to the respective AMP-conjugate (AMP = adenosine monophosphate). This reaction is catalyzed by the A domain, consuming one equivalent of ATP (adenosine triphosphate) in presence of magnesium ions. Activated amino acids are then transferred to the adjacent PCP domain by forming a thioester with the terminal thiol of a phosphopantetheine (ppant) moiety. The amino acid building block is thereby attached to the flexible ppant arm, which directs the amino acid toward the condensation domain (C). At the same time another PCP domain, located downstream within the megasynthase, sequesters a second PCP-bound

amino acid. The condensation domain can now catalyze the amide bond formation between both PCP-bound amino acids (Figure 6).[64,79]

With respect to possible extender units, NRPS systems are more flexible than ribosomal peptide biosynthesis, as adenylation domains can accept a variety of substrates apart from the proteinogenic amino acids. Stachelhaus *et al.* identified ten key amino acid residues responsible for the substrate specificity of A domains. This so-called Stachelhaus code is derived from a crystal structure of an adenylation domain having an amino acid bound to the active site.[80] The side chains that form the cavity around the active site hold responsible for the selectivity of the adenylation domain. This set of side chains is referred to when the Stachelhaus code is mentioned. Using this knowledge, bioinformatics methods have been devised in order to predict an A domain's substrate specificity from its amino acid sequence.[81,82]

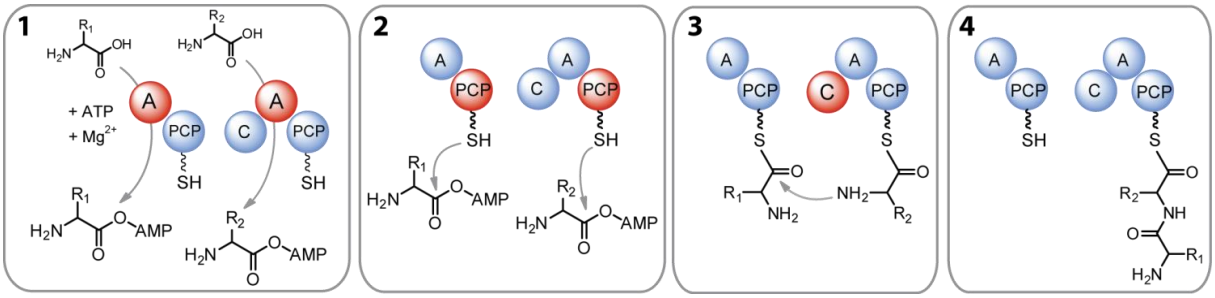


*Figure 6: Principle biosynthetic step leading to peptide bond formation in NRPS. Active domains of each step are highlighted in red. The reaction cycle starts with the activation of an amino acid by an adenylation domain (A) that is specific for the respective amino acid (**1**). A peptidyl carrier protein (PCP) gathers the activated amino acid by transesterification using the terminal thiol of a phosphopantetheine group (**2**). The scheme features a loading unit (A-PCP) which frequently provides the initial amino acid in NRPS clusters.*

In addition to common L-amino acids, many NRPS-derived natural products incorporate unusual building blocks such as D-configured amino acids in their scaffold. Besides the direct activation of D-amino acids by an adenylation domain,[83] epimerization from L to D-configured amino acids is frequently employed to generate D-amino acid moieties in NRPS products.[64] In particular, most D-amino acids are usually not available from primary metabolism, and thus their direct loading appears less favorable. The stereochemical inversion on the assembly line is mediated by dedicated epimerization domains (E) that are integrated into NRPS modules, usually showing the domain order C-A-PCP-E. The proximate module to the E domain is special as its C domain is specific for using a D-configured substrate. For this reason, this C domain is named as $^{D}C_{L}$ domain in contrast to the conventional $^{L}D_{L}$ domain.[84] Incorporation of D-amino acids plays a role for the peptide's metabolic stability, as proteases may not be able to cleave the backbone.[85] Furthermore, the incorporation of non-proteinogenic building blocks contributes to the large structural diversity of NRPS-derived structures, and can contribute to special modes-of-action.[86]

The biosynthetic potential of NRPS is further extended by special C domains (Cy or HC) that are able to cyclize serine, threonine and cysteine to 5-membered heterocycles. These heterocycles can be further

oxidized to form aromatic systems by additional oxidation domains (Ox), which are often found in Cy-containing modules. Modifications like heterocyclization or cross-linking of side chains leads to increased rigidity and stability of NRPS metabolites, and heterocycles are important structural features for chelating metal ions or interacting with nucleotides.[87] Release of PCP-bound intermediates is usually catalyzed by a terminal thioesterase (TE) domain, similar to that found in PKS systems. However, release mechanisms apart from typical TE domains are reported as well.[88] After all, the different types of release mechanisms are the origin of a variety of structural features found in natural product. Depending on whether the NRPS template is released by hydrolysis, intramolecular cyclization, or reductive cleavage the resulting structural feature are lactames, lactones, free acids, imines, or aldehydes.[79,88]

1.3.3 NRPS-PKS Hybrid Pathways

Although NRPS and PKS use different monomeric substrates, their compatible mechanism of chain elongation using ppant-bound intermediates allows the evolution of hybrid systems. PCP and ACP domains have analogous functions but show rather low homology on protein level. Interestingly, their solvent structure is similar with the main difference being the surface charge, which is acidic for ACPs and less polar for PCPs.[79,89] Notwithstanding these differences, true translational hybrids of PKS and NRPS have been known since 1999.[90] The overall diversity of natural products is heavily expanded by this additional option of literally combining NRPS and PKS modules like building blocks. In light of the structures of PKS and NRPS complexes, this possibility is remarkable on first sight as PKS systems have been revealed to act as dimers, whereas the structure of NRPS megasynthases seems to be monomeric.[91] However, a homodimeric NRPS structure has been reported for VibF, a subunit of the vibriobactin cluster.[92] These reports fuel speculation about the true NRPS quaternary structure and raise the question how PKS and NRPS modules can interact in terms of forwarding the PCP/ACP-bound templates at the interface. A recent study gives strong hints that the NRPS part of a NRPS-PKS hybrid is homodimeric, thus allowing a reasonable interaction with the homodimeric PKS part.[93] In particular, the NMR-derived solution structure of the *N*-terminal docking domain from the NRPS module TubC (tubulysin) turned out to be a homodimer featuring a new type of protein fold amongst docking domains. It turned out that there are different types of docking domains between the different combinations of NRPS and PKS systems, which could reflect a separate evolutionary path for these systems.[94]

1.3.4 Deviations from Textbook Biosynthetic Logic

As increasing numbers of PKS and NRPS systems were characterized it became evident that many of these do not follow strict textbook biosynthetic logic, e.g. the co-linearity rule of modular extension.[95,96] Some of these deviations are also found in microsclerodermin biosynthesis which is topic of chapter 3 but also referred to in this chapter. A commonly observed exception is the iterative use of single modules in

PKS assembly lines, as found e.g. in the biosynthesis of stigmatellin or myxochromid.[97,98] Microsclerodermin biosynthesis features two iteratively acting PKS modules as can be seen in Figure 7 (module 1 and 3). There is no hint on how to predict the frequency of iteration. Hence, such iterative type I PKS systems add another level of natural product diversity and complexity to PKS and PKS-NRPS hybrid systems.[99] First direct evidence for an iteratively acting single module is reported for the natural product borellidin.[100] Fortunately, it turned out that iteratively acting PKS modules can be distinguished from modular acting ones by alignment of the protein sequence of KS domains.[101,102] Such in silico predictions help to characterize the modules of a given biosynthetic gene cluster in an early stage. In contrast to the repeated use of a module, module skipping or domain skipping is another observed deviation from textbook logic that was found in both, PKS and NRPS systems.[103,104] In particular, skipping is also observed for the intact catalytic domains MscB and MscD of the microsclerodermin biosynthesis (Figure 7, module 2 and 6). However, it is not evident that this is a real skipping as these intact modules are not translationally fused to the other modules. The exact mechanism on how the ppant-bound templates are forwarded under such conditions is not yet fully understood. One proven mechanism, at least, is the direct transfer from one ppant moiety to another.[105] This is the most obvious and energetically favorable way as the thioester bond persists during the process.

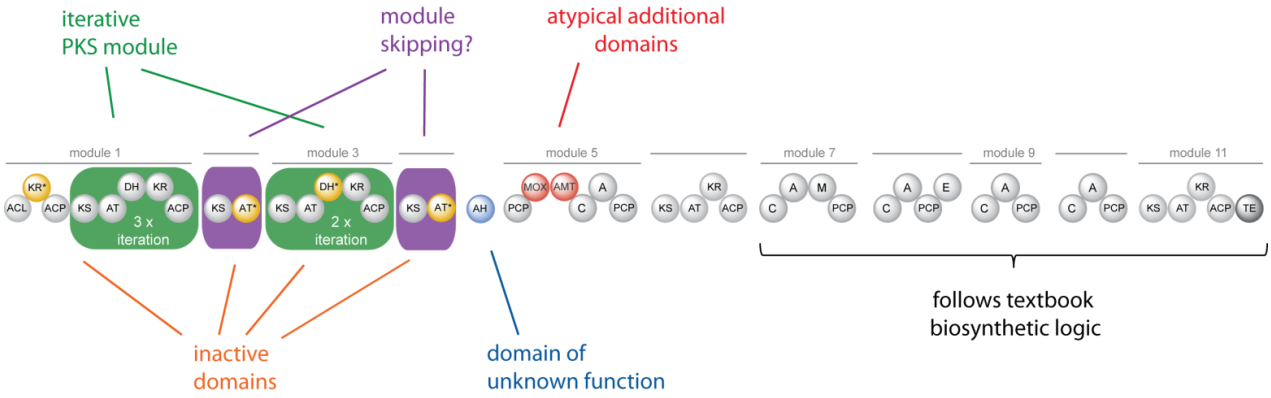


Figure 7: Modules and domains of the biosynthetic gene cluster responsible for microsclerodermin production in S. cellulosum So ce38. The cluster comprises several deviations from textbook biosynthetic logic for NRPS and PKS systems.

Some PKS systems that were identified seemed to be incomplete since they lack an appropriate number of acyl transferases (AT).[106,107] More precisely, there are PKS clusters with just one or two AT domains that hold responsible for furnishing the whole set of modules with monomers. Such clusters are called *trans*-AT cluster as the AT domain can load all KS domains an thereby acts “in trans” in respect to the whole cluster.[108] The expression *cis*-AT clusters, which would comprise typical PKS modules with the complete domain set to perform an elongation step, is rarely used. Evolution of PKS clusters apparently resulted in a PKS architecture (*trans*-AT) that is intrinsically different to the PKS textbook example (*cis*-AT). A phylogenetic analysis of KS domains of *trans*-AT and *cis*-AT clusters results in two distinct clades for both

types, indicating that both systems evolved independently from an ancient FAS system.[109] Numerous clusters were characterized, which ultimately enabled a sequence-based analysis of *trans*-AT related KS domains and permitted the development of a predictive model for KS domain specificity.[110,111]

In addition, some biosynthetic clusters contain additional enzymatic functions such as aminotransferases[112,113], oxidases[114,115], prenyltransferases[116–118], or even enzymes of currently unknown effect, which further increase the variety of chemical entities that can be produced. Moreover, post-assembly line modifications like sulfation or glycosylation were observed among secondary metabolites.[119,120] Most of biosynthetic gene clusters reveal an uncommon architecture and/or functions. In light of this, it is quite obvious that there is still more variation to be observed in the future.

1.4 Analytical Techniques for Natural Product Research

Analytical techniques have always been of importance for natural product research. The advent of liquid chromatography and related separation techniques was a milestone for identification and purification of natural products out of crude biological extracts. Advances in these techniques in combination with the progress in the field of mass spectrometry and DNA sequencing opened up new possibilities of natural product research.[121] Together with the growing number of available genomes and biosynthetic gene clusters, the basis for bioinformatic analysis increases and allows further optimization of the models behind bioinformatics predictions.

1.4.1 Instrumental Analytics in Natural Product Research

Steadily improving techniques in instrumental analytics play a tremendous role in identifying natural products.[122] Modern LC-MS instrumentation has the potential to detect compounds that were missed before, e.g. such that previously evaded detection due to their weak UV absorbance. In line with this, many natural products discovered in the past show a characteristic UV/Vis absorption spectrum; a circumstance based on the fact that UV/Vis detectors were the matter of choice for a long time. Nowadays, mass spectrometric detection is improving the value of chromatographic data by adding highly resolved, accurate *m/z* information that supports the calculation of sum formula.[123] Several types of mass spectrometers can combine wide dynamic range alongside with mass errors of around 1 ppm, e.g. FT-ion traps and modern time-of-flight machines. As a consequence, the problem of coeluting peaks is not that pronounced anymore as high resolution mass spectra allow identification of distinct compounds. At the same time chromatographic separation improved dramatically when stationary phases of the sub-2µm class were introduced to the HPLC market. Highly efficient separations of complex samples became achievable using very short run times. These advances set the stage for reliable chemical screening programs for small molecules, in particular secondary metabolites.

In a similar fashion, progress in proteomics helps to study biosynthetic processes using digested as well as intact proteins of biosynthetic pathways.[124,125] Of special interest amongst the proteomics applications is the ppant ejection assay as introduced by the Kelleher group.[126,127] This methodology opens up the possibility to identify ppant moieties that still harbor biosynthetic intermediates. In particular, the complete ppant moiety is dissociated from an intact protein or from a respective peptide after digestion. This was shown to be applicable very specific by means of infrared multiphoton dissociation (IRMPD). Aside of this important contribution to the field of biosynthesis research, measurement of ACP and PCP domains enabled monitoring of extender unit loading *in vitro*.[128–130] For more aspects of proteomics-linked natural product research the reader is referred to review articles.[124,125]

1.4.2 Genome Mining and in silico Analysis of Biosynthetic Gene Cluster

A milestone in natural product research, especially for microbial secondary metabolites, was the access to whole genome sequences. Versatile high-throughput sequencing techniques recently reached an affordable cost level alongside with high speed and accuracy. The domains of PKS and NRPS clusters have a remarkable sequence similarity on protein level, which facilitates the *in silico* identification. Several bioinformatics tools emerged that take advantage of this circumstance to identify putative secondary metabolite cluster.[131–133] The most convenient tool in this field is likely antiSMASH 2.0 which was made freely available in 2013.[134] It offers automated genome-wide analysis of bacteria, fungi, and archaea for the presence of biosynthetic gene clusters, eventually creating annotated genomes. The genetic potential of an organism is thereby revealed within a short time, which allows numerous ways to proceed towards discovery of the metabolites connected to the predicted pathways, e.g. by mutagenesis, heterologous expression, regulation studies, or by proposals for the putative secondary metabolite structure derived from a given cluster.[36] This knowledge allows focusing on clusters that feature uncommon enzymes or domain organizations as this may be a hint for new chemical moieties. With an organism that is amenable to genetic modifications, such “interesting” clusters should be preferably addressed when thinking about disrupting genes for functional characterization.

Another important aspect deals with prediction of enzyme specificities amongst PKS and NRPS related domains. Biochemical studies and crystal structures of domains led to an understanding of the amino acid residues that govern the specificity or stereochemical course of single biosynthetic steps. A prominent example is the Stachelhaus code which comprises 10 amino acid residues in adenylation (A) domains responsible for the amino acid specificity of the A domain.[80,135] This was further improved by Rausch *et al.* by taking an 8 Å space around the substrate into account.[81] These findings enabled several bioinformatics tools to predict the A domain specificity,[136,137] whereupon the NRPSpredictor2 is most widely used. The substrate prediction is however not equally accurate for the different substrates known so far. While cysteine incorporation is usually predicted without any doubt, prediction of some amino

acids such as tryptophan and lysine relies on an insufficient bioinformatics model.[136] However, the prediction is usually successful to reveal certain substrate classes such as aliphatic side chains with H-bond-donor, aromatic side chains, side chains with hydroxyl groups *et cetera*.

Condensation and ketosynthase domains can be functionally assigned by using the NaPDoS tool in a convenient way.[138] In NRPS systems, different types of C domains have been reported, e.g. C domains incorporating D-amino acids, cyclization domains, epimerization domains, and some mixed mode domains.[81] Among the PKS, Ketosynthase (KS) domains are grouped into *cis*-acting, *trans*-acting, iteratively acting domains. Based solely on genome analysis it is also possible to distinguish a KS domain of a NRPS-PKS interface from a PKS-PKS interface. The specificity of *trans*-AT related KS domains is also matter of recent research.[110,111] Other KS domains that are involved in special types of products are distinguishable as well, i.e. those involved in poly-unsaturated fatty acid and enediyne biosynthesis.[96] Furthermore, the configuration of the hydroxyl group that is created upon reduction of a polyketide extender unit by means of a ketoreduction domain (KR) is also predictable. Caffrey *et al.* proposed two types of KR domains.[139] An A-type KR domain leads to a (*S*)-3-OH-intermediate and the B-type to an (*R*)-3-OH-intermediate bound to the assembly line. The assigned stereochemistry implies that the priority of C2 is higher than of C4. The main discriminant between both types is an conserved LDD motif, which is solely found in B-type domains.[139] This model was further improved by Keatinge-Clay based on the crystal structure for both KR types.[140] Interestingly, the stereogenic center at C3 does have an effect in the next reaction cycle when the DH domain eliminates water. Based on a crystal structure it is proposed that (*S*)-3-OH-intermediates result in *Z*-configured double bonds.[141] Unfortunately, there were not enough DH sequences available at the time of publication to fully prove this. Finally, reduction of the double bond by an enoyl reductase (ER) yields a fully reduced polyketide extender unit. The new stereogenic center, which is set upon enoyl reduction, can be predicted as well.[142]

In conclusion, genome-related techniques are powerful in predicting secondary metabolite clusters together with the putative specificity of single domains and the stereochemistry of structural units. However, it should be clear that results are based on a statistical model derived from all the biosynthetic pathways known so far. This implies that some models are still error-prone as they are based on an underrepresented set of references. On the contrary, other models such as those predicting A-domain specificities are already very good. Fortunately, the models will improve when more biosynthetic pathways are elucidated and published. After all, a full prediction of the compound's structure that is derived from a biosynthetic pathway is not possible as genome mining tools are not able to predict whether a biosynthesis follows the co-linearity rule or whether unforeseen modifications of the compound take place. Hence, in silico techniques to predict secondary metabolites just based on a given genome are still error-prone although yet highly beneficial in the process of biosynthetic cluster characterization.

1.4.3 Secondary Metabolome Mining

Mining the secondary metabolomes of microbial producer benefits from both, the instrumental advances of mass spectrometry, and the genome mining techniques by additionally combining them with bioinformatics tools.[143] Secondary metabolomics is often divided in two fields, namely the targeted and the untargeted approach. A targeted approach aims towards identification of known secondary metabolites, also referred to as dereplication.[144] Dereplication is of high importance when dealing with microbial extracts as chances are that different bacterial strains may produce the same metabolite. Obviously, isolation of already known compounds must be avoided to save time and resources, which are better invested when focusing on new chemical scaffolds. This was a very demoralizing insight in actinomycetes research where thousands of screened strains turned out to produce the same compounds and a plethora of strains had to be screened to finally find some new secondary metabolites.[35]

Aside of a careful selection of the strains that are going to be screened, a proper dereplication process is key to success. A straight forward dereplication strategy is based on a standardized LC-MS screening platform with analytical information available for all known compounds. Samples of unknown composition (e.g. crude extracts prepared from new strains) are compared to the analytical reference like it is commonly done in pesticide analysis or forensics. A typical way to annotate known compounds within such a sample is based on retention time, accurate *m/z*, isotope pattern fit, and occasionally MS/MS data. The comparison of fragment spectra adds another level of confidence. A reliable dereplication is advantageous when a screening project aims towards identification of good producer strains for a distinct compound, such as reported by Krug *et al*. for a set of 98 different strains of *Myxococcus xanthus*.[54] The targeted query amongst the strains allowed the comparison of production rates for cittilin A to eventually identify the best producer for further downstream processes.

When searching for unknown compounds, the situation is more complicated. Especially when dealing with very complex crude extracts it is laborious to determine which signals in a LC-MS chromatogram are related to interesting compounds and which are simply components of the complex nutrition broth. With decreasing relative intensity of a signal of interest, identification of the respective signal gets more complicated. To overcome this problem, statistical tools such as multivariate analysis can be used to highlight differences between distinct sample groups. An interesting application of principal component analysis (PCA) is the analysis of knock-out mutants. After disrupting a biosynthetic gene cluster that is not linked to a known metabolite, it is easier to track the difference between a knock-out mutant and a wild type sample by statistical evaluation rather than by manual inspection.[145,146] It was also shown that strain prioritization as well as sorting of strains according to their secondary metabolome is a suitable application for PCA. [147,148] It should be noted here, that such untargeted analyses request a suitable data processing prior to use in downstream tools. This is usually termed as feature extraction,

which is basically the very first step that aims to extract reasonable information out of raw data on the spectral level. In particular, all *m/z* values are checked for a chromatographic intensity change (i.e. resembling a peak) as only those *m/z* values are likely to represent a real compound eluting from the column. A feature is in terms of LC-MS analysis defined as the triad [*m/z*, retention time, intensity]. It represents the minimal information that is necessary to identify and quantify a compound in a LC-MS measurement. Features from several measurements can be further processed by time alignment or transformed using different mathematical operations.[149–152] Data that is processed this way are subjected to multivariate analysis. In conclusion, multivariate analysis has the potential to highlight differences in secondary metabolome analyses. This is of special importance when changes are not obvious like often experienced for knock-out mutants or differences in regulation.[153]

More specific approaches for the characterization of nonribosomal peptides were devised as well. In particular, the characteristic fragmentation of peptidic bonds enables a *de* novo sequencing strategy for cyclic NRPS. [154,155] However, this holds only true for NRPS compounds that are not highly modified and do not comprise too many structural features apart from NRPS units, i.e. PKS units, alkyl chains, glycosylations *et cetera*. Aside of these *de novo* sequencing techniques, Ibrahim *et al.* developed a library-based screening that is capable of comparing MS/MS fragment spectra of NRPS to known NRPS fragment spectra, thereby adding another way of NRPS dereplication.[156] Another promising approach is linking the genome mining information to mass spectrometric data. Information about a short sequence of amino acids within an unknown compound can facilitate the search for a set of NRPS modules within the respective genome, which encode the proper A-domains.[49] A similar approach is applicable for glycosylations, although the identification of a glycosylation may be not that indicative for a natural product as it is just a post-assembly modification.[157]

In conclusion, mass spectrometric analysis, genome mining, and bioinformatics initiated a new era of natural product discovery by combining the strengths of all methods. However, there are still more problems that need to be addressed and the methods mentioned herein are far from being easily applicable. It became clear that many of these methods rely on MS/MS data. Interestingly, acquisition of MS/MS spectra is still highly biased by the abundance of the secondary metabolites. A fact that is – to the best of my knowledge – not tackled in literature up to now. Chapter 4 of this thesis is basically trying to overcome this drawback by presenting a method that enables targeted MS/MS in an untargeted secondary metabolomics approach.

1.5 Outline of this Work

This thesis comprises several aspects of natural product research. The first part aims toward the discovery and characterization of new secondary metabolites using analytical techniques and methods of genome mining to fully characterize the biosynthetic machinery. Here, the work on pellasoren follows a more traditional “structure-first” approach of isolation and structure elucidation with subsequent characterization of the underlying biosynthesis. The work on microsclerodermin, a previously reported sponge metabolite, features a comprehensive review of this secondary metabolite and its biosynthesis with respect to different sources producing the same compound class thereby extending analysis to various myxobacterial producers. The second part of this thesis deals with the development of analytical methods, including the evaluation and improvement of a LC-MS system that is used for a large screening program together with the development of new methods for untargeted secondary metabolomics.

The myxobacterial strain *S. cellulosum* So ce38 showed promising antimicrobial and cytotoxic activity which led to an in-depth analysis of its metabolite profile in order to find the new secondary metabolites responsible for this bioactivity. In the course of this screening a high abundant peak was identified as pellasoren, a myxobacterial compound previously reported from another *S. cellulosum* strain but at that time not fully characterized. Full structure elucidation of pellasoren was achieved by NMR and MS/MS analyses and the first insight into its biosynthesis was based on bioinformatic analyses of the So ce38 genome. The assigned gene cluster was finally confirmed by creating a knock-out mutant which resulted in abolishment of pellasoren production. Pellasoren biosynthesis incorporates the rare extender unit methoxy-malonyl-CoA that gives rise to an enol ether moiety in the final structure. Interestingly, two slightly different pathways have been proposed for this precursor biogenesis and the genes involved in pellasoren biosynthesis combine aspects of both pathways. A detailed analysis of the biosynthetic domains predicted substrate specificities and stereochemical outcomes of each biosynthetic step. At the same time the total synthesis of this compound was accomplished in the group of Prof. Kalesse (Hannover, Germany), revealing that the *in silico* prediction for one stereochemical center was incorrect. This discrepancy underlined the importance of interdisciplinary approaches in natural product research.

The work on microsclerodermin took advantage of a large-scale screening platform which allowed the analysis of approximately 2000 myxobacterial extracts aimed at identifying distinct compounds. Using this technology, the microsclerodermin compound class was identified in 16 different myxobacterial strains. Moreover, it was shown that the marine-derived metabolite microsclerodermin D is also produced by four different myxobacteria. Thus, microsclerodermin is one of the few examples where an identical natural product is derived from a marine and a terrestrial producer. The finding fueled speculation on whether production in the marine habitat is related to myxobacteria that exist in symbiosis with the sponges. This notion is in agreement with recent results from 16s rRNA analyses of sponge holobioms. In the course of this work, two groups of microsclerodermin producing myxobacteria

were identified comprising 12 and 4 members, respectively. Derivatives of both groups feature an identical core structure but differ in some structural moieties such as hydroxylation, methoxylation, halogenation, or overall length of the side chain. Whole genome sequencing of a representative of both groups provided access to two biosynthetic gene clusters. As a result, group-specific variations of the derivatives were linked to different tailoring enzymes found in both gene clusters.

An additional major part of this work deals with an improved workflow for untargeted secondary metabolomics in natural product research. In particular, development was focused on a more reasonable selection of precursor ions for subsequent acquisition of fragment spectra. The latter is of interest as current developments of computational tools critically rely on fragment spectra for the classification of compounds based on calculated spectra similarities. In the approach taken here, statistical methods are used to create a precursor list reflecting the signals related to bacterial metabolism, thereby increasing the content of LC-MS/MS runs by forwarding only signals of potential interest to MS/MS analysis. This is contrary to classical auto-MS^2 approaches where precursors are usually selected based on signal intensity. As complex biological samples do frequently cover many signals, the most intense signals do not necessarily represent signals of interest. Hence, such a shot gun-like data acquisition is prone to a bias toward unwanted MS/MS scans while important signals may be missed. In line with this, the method described here achieves improved MS/MS scan coverage for precursor ions of low abundance that are usually not covered by an auto-MS^2 experiment. Performance evaluation and assessment of the above method including the effect of ion suppression in crude myxobacterial extracts are examined in order to quantify the effect of matrix components on the detection of known natural products in complex samples. The newly established analytical method was successfully applied to identify the lipothiazole compound class in *S. cellulosum* So ceGT47.

1.6 References

(1) Aminov, R. I.: A brief history of the antibiotic era: lessons learned and challenges for the future. *Front. Microbiol.* **2010**, 1, 134, DOI: 10.3389/fmicb.2010.00134

(2) Carter, K. C.: Koch's postulates in relation to the work of Jacob Henle and Edwin Klebs. *Med. Hist.* **2012**, 29, 353–374, DOI: 10.1017/S0025727300044689

(3) Henle, J.: *Pathologische Untersuchungen*. (Hirschwald, **1840**). VI, 274 S.,

(4) Davies, J.: Are antibiotics naturally antibiotics? *J. Ind. Microbiol. Biotechnol.* **2006**, 33, 496–9, DOI: 10.1007/s10295-006-0112-5

(5) Newman, D. J. & Cragg, G. M.: Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **2012**, 75, 311–35, DOI: 10.1021/np200906s

(6) Appelbaum, P. C.: 2012 and beyond: potential for the start of a second pre-antibiotic era? *J. Antimicrob. Chemother.* **2012**, 67, 2062–8, DOI: 10.1093/jac/dks213

(7) Falagas, M. E. & Bliziotis, I. A.: Pandrug-resistant Gram-negative bacteria: the dawn of the post-antibiotic era? *Int. J. Antimicrob. Agents* **2007**, 29, 630–6, DOI: 10.1016/j.ijantimicag.2006.12.012

(8) Boucher, H. W. *et al.*: Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **2009**, 48, 1–12, DOI: 10.1086/595011

(9) Falagas, M. E. & Kasiakou, S. K.: Colistin: the revival of polymyxins for the management of multidrug-resistant gram-negative bacterial infections. *Clin. Infect. Dis.* **2005**, 40, 1333–41, DOI: 10.1086/429323

(10) Weissman, K. J. & Müller, R.: Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.* **2010**, 27, 1276–95, DOI: 10.1039/c001260m

(11) Garcia, R. & Müller, R.: *The Prokaryotes: Deltaproteobacteria and Epsilonproteobacteria* (Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F.) (Springer, **2014**).

(12) Reichenbach, H.: Myxobacteria, producers of novel bioactive substances. *J. Ind. Microbiol. Biotechnol.* **2001**, 27, 149–156, DOI: 10.1038/sj.jim.7000025

(13) Nan, B. & Zusman, D. R.: Uncovering the mystery of gliding motility in the myxobacteria. *Annu. Rev. Genet.* **2011**, 45, 21–39, DOI: 10.1146/annurev-genet-110410-132547

(14) Kaiser, D.: Are Myxobacteria intelligent? *Front. Microbiol.* **2013**, 4, 335, DOI: 10.3389/fmicb.2013.00335

(15) Schneiker, S. *et al.*: Complete genome sequence of the myxobacterium Sorangium cellulosum. *Nat. Biotechnol.* **2007**, 25, 1281–9, DOI: 10.1038/nbt1354

(16) Schäberle, T. F. *et al.*: Marine myxobacteria as a source of antibiotics--comparison of physiology, polyketide-type genes and antibiotic production of three new isolates of Enhygromyxa salina. *Mar. Drugs* **2010**, 8, 2466–79, DOI: 10.3390/md8092466

(17) Wenzel, S. C. & Müller, R.: *Compr. Nat. Prod. II* (Liu, H.-W. & Mander, L.) (Elsevier, **2010**). null, 189–222, DOI: 10.1016/B978-008045382-8.00645-6

(18) Simister, R. L., Deines, P., Botté, E. S., Webster, N. S. & Taylor, M. W.: Sponge-specific clusters revisited: a comprehensive phylogeny of sponge-associated microorganisms. *Environ. Microbiol.* **2012**, 14, 517–24, DOI: 10.1111/j.1462-2920.2011.02664.x

(19) Ravenschlag, K., Sahm, K., Pernthaler, J. & Amann, R.: High bacterial diversity in permanently cold marine sediments. *Appl. Environ. Microbiol.* **1999**, 65, 3982–9,

(20) Konstantinidis, K. T. & Tiedje, J. M.: Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, 101, 3160–5, DOI: 10.1073/pnas.0308653100

(21) Han, K. *et al.*: Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. *Sci. Rep.* **2013**, 3, 2101, DOI: 10.1038/srep02101

(22) Gerth, K., Pradella, S., Perlova, O., Beyer, S. & Müller, R.: Myxobacteria: proficient producers of novel natural products with various biological activities—past and future biotechnological aspects with the focus on the genus Sorangium. *J. Biotechnol.* **2003**, 106, 233–253, DOI: 10.1016/j.jbiotec.2003.07.015

(23) Xiao, Y., Wei, X., Ebright, R. & Wall, D.: Antibiotic production by myxobacteria plays a role in predation. *J. Bacteriol.* **2011**, 193, 4626–33, DOI: 10.1128/JB.05052-11

(24) Fox, N. K., Brenner, S. E. & Chandonia, J.-M.: SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **2014**, 42, D304–9, DOI: 10.1093/nar/gkt1240

(25) Gaitatzis, N., Kunze, B. & Müller, R.: Novel insights into siderophore formation in myxobacteria. *Chembiochem* **2005**, 6, 365–74, DOI: 10.1002/cbic.200400206

(26) Davies, J. & Ryan, K. S.: Introducing the parvome: bioactive compounds in the microbial world. *ACS Chem. Biol.* **2012**, 7, 252–9, DOI: 10.1021/cb200337h

(27) Gerth, K. & Müller, R.: Moderately thermophilic Myxobacteria: novel potential for the production of natural products isolation and characterization. *Environ. Microbiol.* **2005**, 7, 874–80, DOI: 10.1111/j.1462-2920.2005.00761.x

(28) Höfle, G., Steinmetz, H., Gerth, K. & Reichenbach, H.: Antibiotics from gliding bacteria, XLIV. Ambruticins VS: New members of the antifungal ambruticin family fromSorangium cellulosum. *Liebigs Ann. der Chemie* **1991**, 1991, 941–945, DOI: 10.1002/jlac.1991199101161

(29) Gerth, K., Bedorf, N., Höfle, G., Irschik, H. & Reichenbach, H.: Epothilons A and B: antifungal and cytotoxic compounds from Sorangium cellulosum (Myxobacteria). Production, physico-chemical and biological properties. *J. Antibiot. (Tokyo).* **1996**, 49, 560–3,

(30) Irschik, H., Augustiniak, H., Gerth, K., Höfle, G. & Reichenbach, H.: The ripostatins, novel inhibitors of eubacterial RNA polymerase isolated from myxobacteria. *J. Antibiot. (Tokyo).* **1995**, 48, 787–92,

(31) Irschik, H., Jansen, R., Gerth, K., Höfle, G. & Reichenbach, H.: The sorangicins, novel and powerful inhibitors of eubacterial RNA polymerase isolated from myxobacteria. *J. Antibiot. (Tokyo).* **1987**, 40, 7–13, DOI: 3104268

(32) Gerth, K., Bedorf, N., Irschik, H., Höfle, G. & Reichenbach, H.: The soraphens: a family of novel antifungal compounds from Sorangium cellulosum (Myxobacteria). I. Soraphen A1 alpha: fermentation, isolation, biological properties. *J. Antibiot. (Tokyo).* **1994**, 47, 23–31,

(33) Irschik, H., Reichenbach, H., Höfle, G. & Jansen, R.: The thuggacins, novel antibacterial macrolides from Sorangium cellulosum acting against selected Gram-positive bacteria. *J. Antibiot. (Tokyo).* **2007**, 60, 733–8, DOI: 10.1038/ja.2007.95

(34) Arnison, P. G. *et al.*: Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **2013**, 30, 108–60, DOI: 10.1039/c2np20085f

(35) Müller, R. & Wink, J.: Future potential for anti-infectives from bacteria - How to exploit biodiversity and genomic potential. *Int. J. Med. Microbiol.* **2013**, DOI: 10.1016/j.ijmm.2013.09.004

(36) Wenzel, S. C. & Müller, R.: The impact of genomics on the exploitation of the myxobacterial secondary metabolome. *Nat. Prod. Rep.* **2009**, 26, 1385–407, DOI: 10.1039/b817073h

(37) Schley, C., Altmeyer, M. O., Swart, R., Müller, R. & Huber, C. G.: Proteome analysis of Myxococcus xanthus by off-line two-dimensional chromatographic separation using monolithic poly-(styrene-divinylbenzene) columns combined with ion-trap tandem mass spectrometry. *J. Proteome Res.* **2006**, 5, 2760–8, DOI: 10.1021/pr0602489

(38) Hüttel, S. & Müller, R.: Methods to optimize myxobacterial fermentations using off-gas analysis. *Microb. Cell Fact.* **2012**, 11, 59, DOI: 10.1186/1475-2859-11-59

(39) Traxler, M. F., Watrous, J. D., Alexandrov, T., Dorrestein, P. C. & Kolter, R.: Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome. *MBio* **2013**, 4, DOI: 10.1128/mBio.00459-13

(40) Ueda, K. *et al.*: Wide distribution of interspecific stimulatory events on antibiotic production and sporulation among Streptomyces species. *J. Antibiot. (Tokyo).* **2000**, 53, 979–82, DOI: 11099234

(41) Goldman, B. S. *et al.*: Evolution of sensory complexity recorded in a myxobacterial genome. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, 103, 15200–5, DOI: 10.1073/pnas.0607335103

(42) Huntley, S. *et al.*: Comparative genomic analysis of fruiting body formation in Myxococcales. *Mol. Biol. Evol.* **2011**, 28, 1083–97, DOI: 10.1093/molbev/msq292

(43) Beyer, S., Kunze, B., Silakowski, B. & Müller, R.: Metabolic diversity in myxobacteria: identification of the myxalamid and the stigmatellin biosynthetic gene cluster of Stigmatella aurantiaca Sg a15 and a combined polyketide-(poly)peptide gene cluster from the epothilone producing strain Sorangium cellul. *Biochim. Biophys. Acta - Gene Struct. Expr.* **1999**, 1445, 185–195, DOI: 10.1016/S0167-4781(99)00041-X

(44) Julien, B. & Fehd, R.: Development of a mariner-Based Transposon for Use in Sorangium cellulosum. *Appl. Environ. Microbiol.* **2003**, 69, 6299–6301, DOI: 10.1128/AEM.69.10.6299-6301.2003

(45) Kopp, M. *et al.*: Critical variations of conjugational DNA transfer into secondary metabolite multiproducing Sorangium cellulosum strains So ce12 and So ce56: development of a mariner-based transposon mutagenesis system. *J. Biotechnol.* **2004**, 107, 29–40, DOI: 10.1016/j.jbiotec.2003.09.013

(46) Cortina, N. S., Krug, D., Plaza, A., Revermann, O. & Müller, R.: Myxoprincomide: a natural product from Myxococcus xanthus discovered by comprehensive analysis of the secondary metabolome. *Angew. Chem. Int. Ed. Engl.* **2012**, 51, 811–6, DOI: 10.1002/anie.201106305

(47) Young, J. *et al.*: Elucidation of Gephyronic Acid Biosynthetic Pathway Revealed Unexpected SAM-Dependent Methylations. *J. Nat. Prod.* **2013**, DOI: 10.1021/np400629v

(48) Cortina, N. S., Revermann, O., Krug, D. & Müller, R.: Identification and characterization of the althiomycin biosynthetic gene cluster in Myxococcus xanthus DK897. *Chembiochem* **2011**, 12, 1411–6, DOI: 10.1002/cbic.201100154

(49) Kersten, R. D. *et al.*: A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **2011**, 7, 794–802, DOI: 10.1038/nchembio.684

(50) Gross, F. *et al.*: Bacterial type III polyketide synthases: phylogenetic analysis and potential for the production of novel secondary metabolites by heterologous expression in pseudomonads. *Arch. Microbiol.* **2006**, 185, 28–38, DOI: 10.1007/s00203-005-0059-3

(51) Julien, B. & Shah, S.: Heterologous Expression of Epothilone Biosynthetic Genes in Myxococcus xanthus. *Antimicrob. Agents Chemother.* **2002**, 46, 2772–2778, DOI: 10.1128/AAC.46.9.2772-2778.2002

(52) Fu, J. *et al.*: Efficient transfer of two large secondary metabolite pathway gene clusters into heterologous hosts by transposition. *Nucleic Acids Res.* **2008**, 36, e113, DOI: 10.1093/nar/gkn499

(53) Fu, J. *et al.*: Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nat. Biotechnol.* **2012**, 30, 440–6, DOI: 10.1038/nbt.2183

(54) Krug, D. *et al.*: Discovering the hidden secondary metabolome of Myxococcus xanthus: a study of intraspecific diversity. *Appl. Environ. Microbiol.* **2008**, 74, 3058–68, DOI: 10.1128/AEM.02863-07

(55) Stoops, J. K., Wakil, S. J., Uberbacher, E. C. & Bunick, G. J.: Small-angle neutron-scattering and electron microscope studies of the chicken liver fatty acid synthase. *J. Biol. Chem.* **1987**, 262, 10246–51,

(56) Crisp, J. D. & Wakil, S. J.: Chicken liver fatty acid synthetase: Proteolysis by subtilisin and isolation and properties of palmitoyl thioesterase. *J. Protein Chem.* **1982**, 1, 241–255, DOI: 10.1007/BF01025002

(57) Holak, T. A., Kearsley, S. K., Kim, Y. & Prestegard, J. H.: Three-dimensional structure of acyl carrier protein determined by NMR pseudoenergy and distance geometry calculations. *Biochemistry* **1988**, 27, 6135–6142, DOI: 10.1021/bi00416a046

(58) Staunton, J. & Weissman, K. J.: Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **2001**, 18, 380–416, DOI: 10.1039/a909079g

(59) Birch, A. & Donovan, F.: Studies in relation to Biosynthesis. I. Some possible routes to derivatives of Orcinol and Phloroglucinol. *Aust. J. Chem.* **1953**, 6, 360, DOI: 10.1071/CH9530360

(60) Donadio, S., Staver, M. J., McAlpine, J. B., Swanson, S. J. & Katz, L.: Modular organization of genes required for complex polyketide biosynthesis. *Science* **1991**, 252, 675–9,

(61) Hopwood, D. a.: Genetic Contributions to Understanding Polyketide Synthases. *Chem. Rev.* **1997**, 97, 2465–2498,

(62) Staunton, J. *et al.*: Evidence for a double-helical structure for modular polyketide synthases. *Nat. Struct. Biol.* **1996**, 3, 188–192, DOI: 10.1038/nsb0296-188

(63) Kao, C. M., Pieper, R., Cane, D. E. & Khosla, C.: Evidence for two catalytically independent clusters of active sites in a functional modular polyketide synthase. *Biochemistry* **1996**, 35, 12363–8, DOI: 10.1021/bi9616312

(64) Fischbach, M. a & Walsh, C. T.: Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **2006**, 106, 3468–96, DOI: 10.1021/cr0503097

(65) Crosby, J. & Crump, M. P.: The structural role of the carrier protein--active controller or passive carrier. *Nat. Prod. Rep.* **2012**, 29, 1111–37, DOI: 10.1039/c2np20062g

(66) Keatinge-Clay, A. T.: The structures of type I polyketide synthases. *Nat. Prod. Rep.* **2012**, 29, 1050–73, DOI: 10.1039/c2np20019h

(67) Wenzel, S. C. *et al.*: On the biosynthetic origin of methoxymalonyl-acyl carrier protein, the substrate for incorporation of “glycolate” units into ansamitocin and soraphen A. *J. Am. Chem. Soc.* **2006**, 128, 14325–36, DOI: 10.1021/ja064408t

(68) Carroll, B. J. *et al.*: Identification of a set of genes involved in the formation of the substrate for the incorporation of the unusual “glycolate” chain extension unit in ansamitocin biosynthesis. *J. Am. Chem. Soc.* **2002**, 124, 4176–7,

(69) Wu, K., Chung, L., Revill, W. P., Katz, L. & Reeves, C. D.: The FK520 gene cluster of Streptomyces hygroscopicus var. ascomyceticus (ATCC 14891) contains genes for biosynthesis of unusual polyketide extender units. *Gene* **2000**, 251, 81–90,

(70) Chan, Y. a *et al.*: Hydroxymalonyl-acyl carrier protein (ACP) and aminomalonyl-ACP are two additional type I polyketide synthase extender units. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, 103, 14349–54, DOI: 10.1073/pnas.0603748103

(71) Choi, S.-S., Hur, Y.-A., Sherman, D. H. & Kim, E.-S.: Isolation of the biosynthetic gene cluster for tautomycetin, a linear polyketide T cell-specific immunomodulator from Streptomyces sp. CK4412. *Microbiology* **2007**, 153, 1095–102, DOI: 10.1099/mic.0.2006/003194-0

(72) Moore, B. S. & Hertweck, C.: Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Nat. Prod. Rep.* **2002**, 19, 70–99, DOI: 10.1039/b003939j

(73) Yadav, G., Gokhale, R. S. & Mohanty, D.: Computational Approach for Prediction of Domain Organization and Substrate Specificity of Modular Polyketide Synthases. *J. Mol. Biol.* **2003**, 328, 335–363, DOI: 10.1016/S0022-2836(03)00232-8

(74) Long, P. F. *et al.*: Engineering specificity of starter unit selection by the erythromycin-producing polyketide synthase. *Mol. Microbiol.* **2002**, 43, 1215–25,

(75) Dunn, B. J. & Khosla, C.: Engineering the acyltransferase substrate specificity of assembly line polyketide synthases. *J. R. Soc. Interface* **2013**, 10, 20130297, DOI: 10.1098/rsif.2013.0297

(76) Weissman, K. J.: Introduction to polyketide biosynthesis. *Methods Enzymol.* **2009**, 459, 3–16, DOI: 10.1016/S0076-6879(09)04601-1

(77) Müller, R.: Don't classify polyketide synthases. *Chem. Biol.* **2004**, 11, 4–6, DOI: 10.1016/j.chembiol.2004.01.005

(78) Lipmann, F., Gevers, W., Kleinkauf, H. & Roskoski, R.: Polypeptide synthesis on protein templates: the enzymatic synthesis of gramicidin S and tyrocidine. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1971**, 35, 1–34, DOI: 10.1002/9780470122808

(79) Sieber, S. a & Marahiel, M. a: Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. *Chem. Rev.* **2005**, 105, 715–38, DOI: 10.1021/cr0301191

(80) Stachelhaus, T., Mootz, H. D. & Marahiel, M. A.: The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **1999**, 6, 493–505, DOI: 10.1016/S1074-5521(99)80082-9

(81) Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D. H.: Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **2005**, 33, 5799–808, DOI: 10.1093/nar/gki885

(82) Röttig, M. *et al.*: NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **2011**, 39, W362–7, DOI: 10.1093/nar/gkr323

(83) Hoffmann, K., Schneider-Scherzer, E., Kleinkauf, H. & Zocher, R.: Purification and characterization of eucaryotic alanine racemase acting as key enzyme in cyclosporin biosynthesis. *J. Biol. Chem.* **1994**, 269, 12710–4,

(84) Linne, U., Doekel, S. & Marahiel, M. A.: Portability of Epimerization Domain and Role of Peptidyl Carrier Protein on Epimerization Activity in Nonribosomal Peptide Synthetases †. *Biochemistry* **2001**, 40, 15824–15834, DOI: 10.1021/bi011595t

(85) Hamamoto, K., Kida, Y., Zhang, Y., Shimizu, T. & Kuwano, K.: Antimicrobial activity and stability to proteolysis of small linear cationic peptides with D-amino acid substitutions. *Microbiol. Immunol.* **2002**, 46, 741–9,

(86) Knerr, P. J. *et al.*: Non-proteinogenic amino acids in lacticin 481 analogues result in more potent inhibition of peptidoglycan transglycosylation. *ACS Chem. Biol.* **2012**, 7, 1791–5, DOI: 10.1021/cb300372b

(87) Roy, R. S., Gehring, A. M., Milne, J. C., Belshaw, P. J. & Walsh, C. T.: Thiazole and oxazole peptides: biosynthesis and molecular machinery. *Nat. Prod. Rep.* **1999**, 16, 249–263, DOI: 10.1039/a806930a

(88) Du, L. & Lou, L.: PKS and NRPS release mechanisms. *Nat. Prod. Rep.* **2010**, 27, 255–78, DOI: 10.1039/b912037h

(89) Weber, T., Baumgartner, R., Renner, C., Marahiel, M. A. & Holak, T. A.: Solution structure of PCP, a prototype for the peptidyl carrier domains of modular peptide synthetases. *Structure* **2000**, 8, 407–418, DOI: 10.1016/S0969-2126(00)00120-9

(90) Paitan, Y., Alon, G., Orr, E., Ron, E. Z. & Rosenberg, E.: The first gene in the biosynthesis of the polyketide antibiotic TA of Myxococcus xanthus codes for a unique PKS module coupled to a peptide synthetase. *J. Mol. Biol.* **1999**, 286, 465–74, DOI: 10.1006/jmbi.1998.2478

(91) Sieber, S. A. *et al.*: Evidence for a Monomeric Structure of Nonribosomal Peptide Synthetases. *Chem. Biol.* **2002**, 9, 997–1008, DOI: 10.1016/S1074-5521(02)00214-4

(92) Hillson, N. J. & Walsh, C. T.: Dimeric structure of the six-domain VibF subunit of vibriobactin synthetase: mutant domain activity regain and ultracentrifugation studies. *Biochemistry* **2003**, 42, 766–75, DOI: 10.1021/bi026903h

(93) Richter, C. D., Nietlispach, D., Broadhurst, R. W. & Weissman, K. J.: Multienzyme docking in hybrid megasynthetases. *Nat. Chem. Biol.* **2008**, 4, 75–81, DOI: 10.1038/nchembio.2007.61

(94) Weissman, K. J. & Müller, R.: Protein-protein interactions in multienzyme megasynthetases. *Chembiochem* **2008**, 9, 826–48, DOI: 10.1002/cbic.200700751

(95) Moss, S. J., Martin, C. J. & Wilkinson, B.: Loss of co-linearity by modular polyketide synthases: a mechanism for the evolution of chemical diversity. *Nat. Prod. Rep.* **2004**, 21, 575–93, DOI: 10.1039/b315020h

(96) Hertweck, C.: The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed. Engl.* **2009**, 48, 4688–716, DOI: 10.1002/anie.200806121

(97) Gaitatzis, N. *et al.*: The biosynthesis of the aromatic myxobacterial electron transport inhibitor stigmatellin is directed by a novel type of modular polyketide synthase. *J. Biol. Chem.* **2002**, 277, 13082–90, DOI: 10.1074/jbc.M111738200

(98) Wenzel, S. C. *et al.*: Structure and biosynthesis of myxochromides S1-3 in Stigmatella aurantiaca: evidence for an iterative bacterial type I polyketide synthase and for module skipping in nonribosomal peptide biosynthesis. *Chembiochem* **2005**, 6, 375–85, DOI: 10.1002/cbic.200400282

(99) Fisch, K. M.: Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS. *RSC Adv.* **2013**, 3, 18228, DOI: 10.1039/c3ra42661k

(100) Olano, C. *et al.*: Evidence from engineered gene fusions for the repeated use of a module in a modular polyketide synthase. *Chem. Commun.* **2003**, 2780, DOI: 10.1039/b310648a

(101) Yadav, G., Gokhale, R. S. & Mohanty, D.: Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput. Biol.* **2009**, 5, e1000351, DOI: 10.1371/journal.pcbi.1000351

(102) Ziemert, N. *et al.*: The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* **2012**, 7, e34064, DOI: 10.1371/journal.pone.0034064

(103) Gaitatzis, N., Hans, a, Müller, R. & Beyer, S.: The mtaA gene of the myxothiazol biosynthetic gene cluster from Stigmatella aurantiaca DW4/3-1 encodes a phosphopantetheinyl transferase that activates polyketide synthases and polypeptide synthetases. *J. Biochem.* **2001**, 129, 119–24,

(104) Beck, B. J., Yoon, Y. J., Reynolds, K. A. & Sherman, D. H.: The Hidden Steps of Domain Skipping. *Chem. Biol.* **2002**, 9, 575–583, DOI: 10.1016/S1074-5521(02)00146-1

(105) Thomas, I., Martin, C. J., Wilkinson, C. J., Staunton, J. & Leadlay, P. F.: Skipping in a Hybrid Polyketide Synthase. *Chem. Biol.* **2002**, 9, 781–787, DOI: 10.1016/S1074-5521(02)00164-3

(106) Moldenhauer, J., Chen, X.-H., Borriss, R. & Piel, J.: Biosynthesis of the antibiotic bacillaene, the product of a giant polyketide synthase complex of the trans-AT family. *Angew. Chem. Int. Ed. Engl.* **2007**, 46, 8195–7, DOI: 10.1002/anie.200703386

(107) Piel, J.: A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, 99, 14002–7, DOI: 10.1073/pnas.222481399

(108) Piel, J.: Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat. Prod. Rep.* **2010**, 27, 996–1047, DOI: 10.1039/b816430b

(109) Piel, J., Hui, D., Fusetani, N. & Matsunaga, S.: Targeting modular polyketide synthases with iteratively acting acyltransferases from metagenomes of uncultured bacterial consortia. *Environ. Microbiol.* **2004**, 6, 921–7, DOI: 10.1111/j.1462-2920.2004.00531.x

(110) Jenner, M. *et al.*: Substrate specificity in ketosynthase domains from trans-AT polyketide synthases. *Angew. Chem. Int. Ed. Engl.* **2013**, 52, 1143–7, DOI: 10.1002/anie.201207690

(111) Nguyen, T. *et al.*: Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol.* **2008**, 26, 225–33, DOI: 10.1038/nbt1379

(112) Aron, Z. D., Dorrestein, P. C., Blackhall, J. R., Kelleher, N. L. & Walsh, C. T.: Characterization of a new tailoring domain in polyketide biogenesis: the amine transferase domain of MycA in the mycosubtilin gene cluster. *J. Am. Chem. Soc.* **2005**, 127, 14986–7, DOI: 10.1021/ja055247g

(113) Tillett, D. *et al.*: Structural organization of microcystin biosynthesis in Microcystis aeruginosa PCC7806: an integrated peptide-polyketide synthetase system. *Chem. Biol.* **2000**, 7, 753–64,

(114) Viehrig, K. *et al.*: Concerted Action of P450 Plus Helper Protein To Form the Amino-hydroxy-piperidone Moiety of the Potent Protease Inhibitor Crocapeptin. *J. Am. Chem. Soc.* **2013**, 135, 16885–94, DOI: 10.1021/ja4047153

(115) Zerbe, K. *et al.*: Crystal structure of OxyB, a cytochrome P450 implicated in an oxidative phenol coupling reaction during vancomycin biosynthesis. *J. Biol. Chem.* **2002**, 277, 47476–85, DOI: 10.1074/jbc.M206342200

(116) Edwards, D. J. & Gerwick, W. H.: Lyngbyatoxin biosynthesis: sequence of biosynthetic gene cluster and identification of a novel aromatic prenyltransferase. *J. Am. Chem. Soc.* **2004**, 126, 11432–3, DOI: 10.1021/ja047876g

(117) Calderone, C. T., Kowtoniuk, W. E., Kelleher, N. L., Walsh, C. T. & Dorrestein, P. C.: Convergence of isoprene and polyketide biosynthetic machinery: isoprenyl-S-carrier proteins in the pksX pathway of Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, 103, 8977–82, DOI: 10.1073/pnas.0603148103

(118) Kopp, M. *et al.*: Insights into the complex biosynthesis of the leupyrrins in Sorangium cellulosum So ce690. *Mol. Biosyst.* **2011**, 7, 1549–63, DOI: 10.1039/c0mb00240b

(119) Zander, W. *et al.*: Sulfangolids, macrolide sulfate esters from Sorangium cellulosum. *Chemistry* **2012**, 18, 6264–71, DOI: 10.1002/chem.201100851

(120) Walsh, C., Freel Meyers, C. L. & Losey, H. C.: Antibiotic glycosyltransferases: antibiotic maturation and prospects for reprogramming. *J. Med. Chem.* **2003**, 46, 3425–36, DOI: 10.1021/jm030257i

(121) Bode, H. B. & Müller, R.: The impact of bacterial genomics on natural product research. *Angew. Chem. Int. Ed. Engl.* **2005**, 44, 6828–46, DOI: 10.1002/anie.200501080

(122) Kuehnbaum, N. L. & Britz-McKibbin, P.: New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. *Chem. Rev.* **2013**, 113, 2437–68, DOI: 10.1021/cr300484s

(123) Kind, T. & Fiehn, O.: Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **2007**, 8, 105, DOI: 10.1186/1471-2105-8-105

(124) Dorrestein, P. C. & Kelleher, N. L.: Dissecting non-ribosomal and polyketide biosynthetic machineries using electrospray ionization Fourier-Transform mass spectrometry. *Nat. Prod. Rep.* **2006**, 23, 893–918, DOI: 10.1039/b511400b

(125) Kersten, R. D., Meehan, M. J. & Dorrestein, P. C.: (Liu, H.-W. (Ben) & Mander, L. B. T.-C. N. P. I. I.) (Elsevier, **2010**). 389–456, DOI: 10.1016/B978-008045382-8.00711-5

(126) Dorrestein, P. C. *et al.*: Facile detection of acyl and peptidyl intermediates on thiotemplate carrier domains via phosphopantetheinyl elimination reactions during tandem mass spectrometry. *Biochemistry* **2006**, 45, 12756–66, DOI: 10.1021/bi061169d

(127) Bumpus, S. B. & Kelleher, N. L.: Accessing natural product biosynthetic processes by mass spectrometry. *Curr. Opin. Chem. Biol.* **2008**, 12, 475–82, DOI: 10.1016/j.cbpa.2008.07.022

(128) Kelleher, N. L. & Hicks, L. M.: Contemporary mass spectrometry for the direct detection of enzyme intermediates. *Curr. Opin. Chem. Biol.* **2005**, 9, 424–30, DOI: 10.1016/j.cbpa.2005.08.018

(129) Gatto, G. J., McLoughlin, S. M., Kelleher, N. L. & Walsh, C. T.: Elucidating the substrate specificity and condensation domain activity of FkbP, the FK520 pipecolate-incorporating enzyme. *Biochemistry* **2005**, 44, 5993–6002, DOI: 10.1021/bi050230w

(130) Dorrestein, P. C. *et al.*: Activity screening of carrier domains within nonribosomal peptide synthetases using complex substrate mixtures and large molecule mass spectrometry. *Biochemistry* **2006**, 45, 1537–46, DOI: 10.1021/bi052333k

(131) Bachmann, B. O. & Ravel, J.: *Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol.* (Elsevier Inc., **2009**). 458, 181–217, DOI: 10.1016/S0076-6879(09)04808-3

(132) Weber, T. *et al.*: CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.* **2009**, 140, 13–17, DOI: 10.1016/j.jbiotec.2009.01.007

(133) Anand, S. *et al.*: SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.* **2010**, 38, W487–96, DOI: 10.1093/nar/gkq340

(134) Blin, K. *et al.*: antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* **2013**, 41, W204–12, DOI: 10.1093/nar/gkt449

(135) Challis, G. L., Ravel, J. & Townsend, C. A.: Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **2000**, 7, 211–224, DOI: 10.1016/S1074-5521(00)00091-0

(136) Röttig, M. *et al.*: NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **2011**, 39, W362–7, DOI: 10.1093/nar/gkr323

(137) Prieto, C., García-Estrada, C., Lorenzana, D. & Martín, J. F.: NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics* **2012**, 28, 426–7, DOI: 10.1093/bioinformatics/btr659

(138) Ziemert, N. *et al.*: The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* **2012**, 7, e34064, DOI: 10.1371/journal.pone.0034064

(139) Caffrey, P.: Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. *Chembiochem* **2003**, 4, 654–7, DOI: 10.1002/cbic.200300581

(140) Keatinge-Clay, A. T.: A tylosin ketoreductase reveals how chirality is determined in polyketides. *Chem. Biol.* **2007**, 14, 898–908, DOI: 10.1016/j.chembiol.2007.07.009

(141) Keatinge-Clay, A.: Crystal structure of the erythromycin polyketide synthase dehydratase. *J. Mol. Biol.* **2008**, 384**,** 941–53, DOI: 10.1016/j.jmb.2008.09.084
(142) Kwan, D. H. *et al.*: Prediction and manipulation of the stereochemistry of enoylreduction in modular polyketide synthases. *Chem. Biol.* **2008**, 15**,** 1231–40, DOI: 10.1016/j.chembiol.2008.09.012
(143) Krug, D. & Müller, R.: Secondary metabolomics: the impact of mass spectrometry- based approaches on the discovery and characterization of microbial natural products.
(144) Constant, H. L. & Beecher, C. W. W.: A Method for the Dereplication of Natural Product Extracts using Electrospray HPLC/MS. *Nat. Prod. Lett.* **1995**, 6**,** 193–196, DOI: 10.1080/10575639508043158
(145) Cortina, N. S., Krug, D., Plaza, A., Revermann, O. & Müller, R.: Myxoprincomide: a natural product from Myxococcus xanthus discovered by comprehensive analysis of the secondary metabolome. *Angew. Chem. Int. Ed. Engl.* **2012**, 51**,** 811–6, DOI: 10.1002/anie.201106305
(146) Asamizu, S., Abugreen, M. & Mahmud, T.: Comparative metabolomic analysis of an alternative biosynthetic pathway to pseudosugars in Actinosynnema mirum DSM 43827. *Chembiochem* **2013**, 14**,** 1548–51, DOI: 10.1002/cbic.201300384
(147) Hou, Y. *et al.*: Microbial strain prioritization using metabolomics tools for the discovery of natural products. *Anal. Chem.* **2012**, 84**,** 4277–83, DOI: 10.1021/ac202623g
(148) Farag, M. A., Weigend, M., Luebert, F., Brokamp, G. & Wessjohann, L. A.: Phytochemical, phylogenetic, and anti-inflammatory evaluation of 43 Urtica accessions (stinging nettle) based on UPLC–Q-TOF-MS metabolomic profiles. *Phytochemistry* **2013**, DOI: http://dx.doi.org/10.1016/j.phytochem.2013.09.016
(149) Van den Berg, R. a, Hoefsloot, H. C. J., Westerhuis, J. a, Smilde, A. K. & van der Werf, M. J.: Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006**, 7**,** 142, DOI: 10.1186/1471-2164-7-142
(150) Christin, C. *et al.*: Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal. Chem.* **2008**, 80**,** 7012–21, DOI: 10.1021/ac800920h
(151) Tautenhahn, R. *et al.*: metaXCMS: second-order analysis of untargeted metabolomics data. *Anal. Chem.* **2011**, 83**,** 696–700, DOI: 10.1021/ac102980g
(152) Benton, H. P., Wong, D. M., Trauger, S. a & Siuzdak, G.: XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. *Anal. Chem.* **2008**, 80**,** 6382–9, DOI: 10.1021/ac800795f
(153) Vinayavekhin, N. & Saghatelian, A.: Regulation of alkyl-dihydrothiazole-carboxylates (ATCs) by iron and the pyochelin gene cluster in Pseudomonas aeruginosa. *ACS Chem. Biol.* **2009**, 4**,** 617–23, DOI: 10.1021/cb900075n
(154) Ng, J. *et al.*: Dereplication and de novo sequencing of nonribosomal peptides. *Nat. Methods* **2009**, 6**,** 596–9, DOI: 10.1038/nmeth.1350
(155) Kavan, D., Kuzma, M., Lemr, K., Schug, K. A. & Havlicek, V.: CYCLONE-A Utility for De Novo Sequencing of Microbial Cyclic Peptides. *J. Am. Soc. Mass Spectrom.* **2013**, 24**,** 1177–84, DOI: 10.1007/s13361-013-0652-7
(156) Ibrahim, A. *et al.*: Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, 109**,** 19196–201, DOI: 10.1073/pnas.1206376109
(157) Kersten, R. D. *et al.*: Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, DOI: 10.1073/pnas.1315492110

Chapter 2 – Pellasoren

Pellasoren: Structure Elucidation, Biosynthesis and Total Synthesis of a Cytotoxic Secondary Metabolite from *Sorangium cellulosum*

Christine Jahns[+a], Thomas Hoffmann[+b], Stefan Müller[b], Klaus Gerth[c], Peter Washausen[c], Gerhard Höfle[c], Hans Reichenbach[c], Markus Kalesse[a], and Rolf Müller[b]**

[+]These authors contributed equally to this work

Angewandte Chemie Int. Ed. **2012**, 51, 5239 –5243 (DOI: 10.1002/anie.201200327)

Angewandte Chemie **2012**, 124, 5330 –5334 (DOI: 10.1002/ange.201200327)

Published online: 05.04.2012

2 Pellasoren

2.1 Introduction

Myxobacteria are efficient producers of numerous secondary metabolites, and the genus *Sorangium* is frequently described as an proficient source for new, biologically active natural products.[1-3] We report here the discovery and complete structure elucidation of pellasoren from the myxobacterium *Sorangium cellulosum*. Identification of the corresponding *pel* gene cluster from *S. cellulosum* So ce38 allowed us to establish a model for pellasoren biosynthesis, providing evidence for an unusual route to glycolate extender unit generation. Moreover, we present the first total synthesis of pellasoren and thereby confirm the absolute configuration of this natural product.

2.2 Results and Discussion

Pellasoren (**1**) was initially isolated from *S. cellulosum* So ce35 in the course of an activity-guided discovery program.[4] Additionally, we recently identified **1** in relatively high amounts in extracts from strain *S. cellulosum* So ce38 using LC-MS analysis. Here we determine its cytotoxicity against HCT-116 human colon cancer cells at a concentration of 155 nM (IC50). Full structural elucidation by NMR and ESI-MS analysis was performed and confirmed the identity of pellasoren from both sources (Supporting Figures 1 to 3 and Table 1).

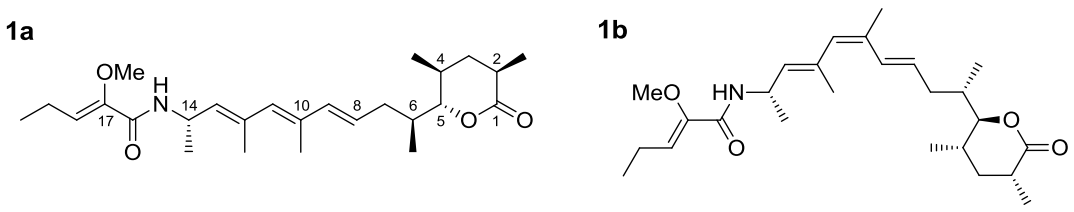


*Figure 1: Structure of pellasoren A (**1a**) and B (**1b**).*

The pellasoren scaffold features an unusual enol ether moiety, also known from a small number of other natural products, which was corroborated by specific HMBC correlations between a methoxy signal and sp2 hybridized carbon atoms (Supporting Figure 1).[5-7] The lactone moiety was identified through HMBC correlation between C1 and C5 and characteristic shifts for H-5 and C5 of 4.02 and 90.3 ppm, respectively. The position of the amide bond was assigned on the basis of indicative fragments in CID spectra. Efforts were made to establish the molecule's relative configuration by using NOE spectroscopy: ROESY interactions together with molecular modelling suggest an *anti*-configuration

of the substituents at C4 and C5, as well as a *syn*-configuration of the methyl groups at C2 and C4 (Supporting Figure 2). The stereogenic centre at C14 maintains the configuration derived from the incorporation of an L-alanine building block during biosynthesis. Stereochemical assignments which could initially not be validated by NOE analysis, such as the configuration at C6, were later established following total synthesis (see below). Additionally, two isomeric pellasorens differing at the C10-C11 double bond configuration were isolated from *S. cellulosum* extracts. Pellasoren A (**1a)** represents the (*E*)-C10-C11 configuration while the B derivative **1b** has (*Z*)-C10-C11 configuration which can be rationalized by isomerization of the corresponding double bond.

Following the structural elucidation of pellasoren, we set out to identify the underlying biosynthetic machinery using fragmentary whole-genome sequence information for strain *S. cellulosum* So ce38. Assuming that pellasoren is most likely the product of a hybrid PKS/NRPS biosynthetic pathway,[8] the full complement of putative PKR/NRPS-related domains encoded in the So ce38 genome was annotated using bioinformatic tools.[9] Using the presumed incorporation of alanine into pellasoren as a guide, we identified a genomic region ~ 57,500 bp in size, containing seven characteristic PKS modules organized as an apparent operon which also encodes one adenylation (A) domain exhibiting the required predicted substrate specificity for alanine (Figure 2a). Insertion by single-crossover homologous recombination of a plasmid conferring hygromycin resistance into module 7 led to the abolishment of pellasoren production and thus ultimately confirmed involvement of the *pel* locus in pellasoren biosynthesis (Supporting Figure 5).

According to common PKS- and NRPS biosynthetic logic, we reasoned that pellasoren formation starts with the condensation between an ACP-bound propionate and a rare glycolate-derived extender unit by module 1, followed by extension with alanine in module 2. Biosynthesis then continues via elongation with two methylmalonyl-CoA units (mM-CoA, modules 3 and 4), one malonyl-CoA (M-CoA, module 5) and another three mM-CoA extender units (modules 6 to 8), and finally the full-length intermediate undergoes lactone ring formation and is released by the terminal TE domain in module 8 (Figure 2a). The order and function of domains encoded by genes *pel*A to *pel*F is pleasingly colinear with the proposed biosynthetic route.

Furthermore, both the predicted substrate specificities of AT and A domains (Supporting Table 5 and Figure 4) and the presence of KR, DH and ER domains necessary to achieve the appropriate reduction of extender units agree well with the pellasoren **1** scaffold. Modules 2 and 7 encode seemingly superfluous ketoreductase (PelA-KR_2) and dehydratase domains (PelE-DH), respectively; however, these domains are likely inactive as judged on the basis of sequence analysis, revealing deviations inside otherwise highly conserved amino acid motifs (Supporting Tables 4 and 6).

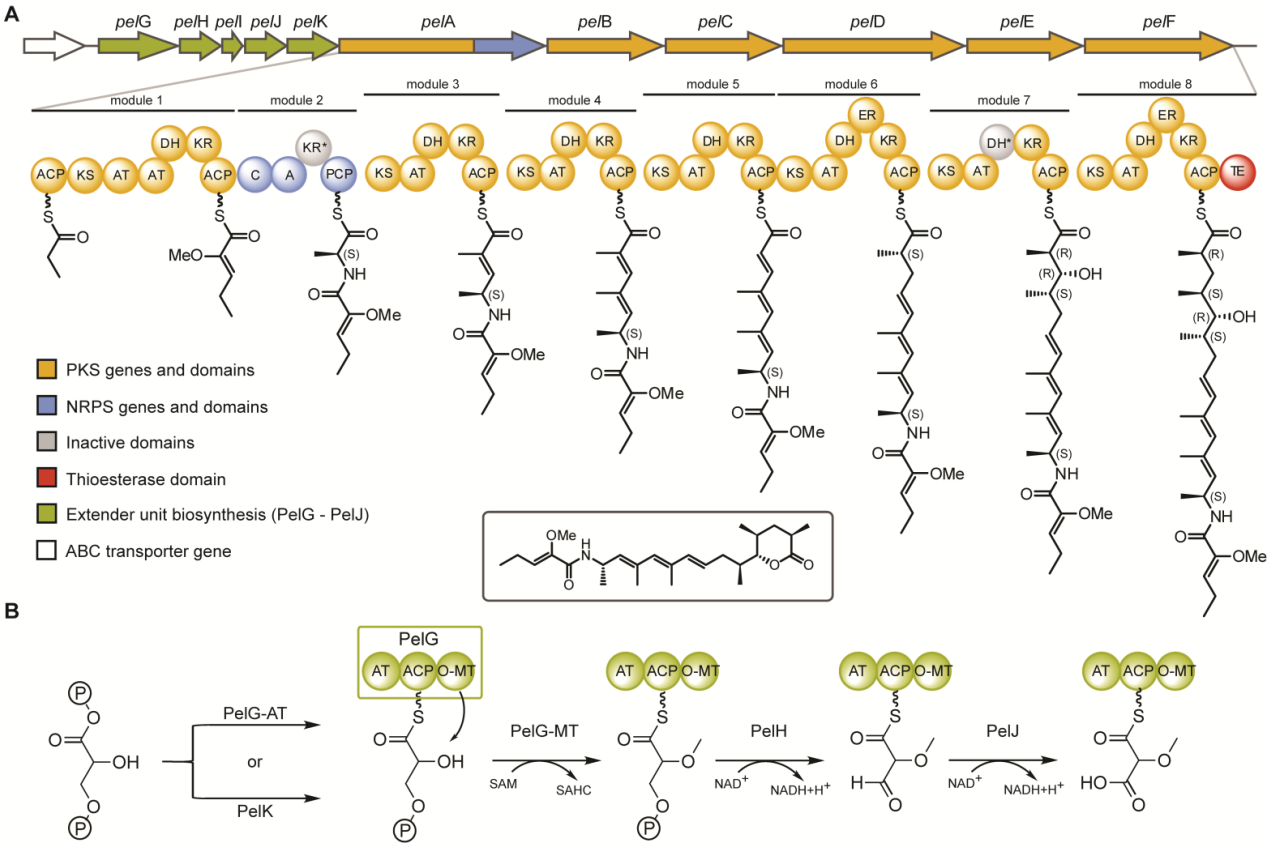


Figure 2: A) The pel biosynthetic gene cluster in Sorangium cellulosum So ce38 and the proposed biosynthetic route to pellasoren. B) Putative biosynthetic pathway to the glycolate extender unit, allowing for alternative loading of phosphoglycerate onto the PelG-ACP by PelK or PelG-AT. A: adenylation domain, ACP: acyl carrier protein, AT: acyltransferase domain, C: condensation domain, DH: dehydratase domain, ER: enoylreductase domain, KR: ketoreductase domain, KS: ketosynthase domain, TE: thioesterase domain.

Pellasoren (**1**) features several chiral centres (C2, C4, C5, C6, C14) of which the configuration is set during its biosynthesis. The absence of an epimerization domain within the NRPS module 2 in PelA indicates an S-configured C14 for the alanine-derived building block. The configuration at C4 and C5 results from extender unit reduction by PelE-KR in module 7, whereas stereocentres at C2 and C6 are generated by the ER domains PelD-ER (module 6) and PelF-ER (module 8, Figure 2a). All active KR domains are of B-type[10] which is in agreement with the 4*R*,5*R* configuration of pellasoren. Kwan *et al.* highlighted several key amino acid residues in ER domains which appear to exert a crucial influence on the stereochemical outcome of the reduction resulting in either 2*R* or 2*S* configuration of the intermediary building block.[11] A tryptophan residue at position 52 has been shown to direct the conformation to 2*S*, whereas absence of Trp52 favours the 2*R* conformation. In pellasoren biosynthesis, both ER domains should produce 2*R* centres since they lack Trp52. Nevertheless, they appear to establish converse stereogenic centres during biosynthesis: 2*S* (by PelD-ER) and 2*R* (by PelF-ER). The conflict between the bioinformatics prediction and the proposed NMR-based configuration of pellasoren (**1**) could only be solved by total synthesis of the NMR-based structure. We therefore decided to approach the synthesis of pellasoren A (**1a**).

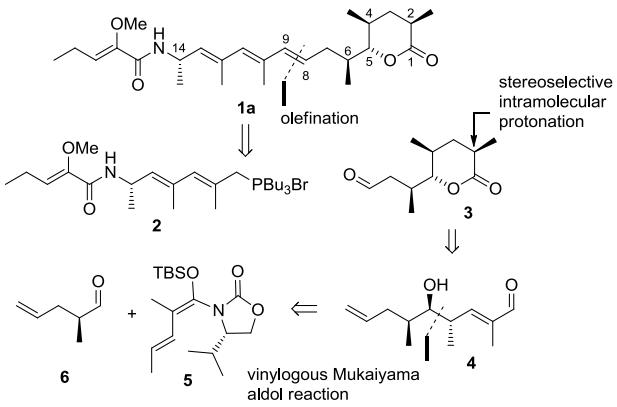


Figure 3: Retrosynthetic analysis of pellasoren A.

Our retrosynthetic analysis divides pellasoren into two key building blocks (**2** and **3**) which are supposed to be connected by an olefination reaction in order to install the sensitive triene at a late stage during the synthesis (Figure 3). In synthetic direction, the western fragment **2** is accessible by a sequence of eight steps, starting from L-alanine (**9**). The eastern fragment **3** contains the lactone subunit and it was planned to be generated by an intramolecular stereoselective protonation of an aldehyde-derived enolate as the key step. This particular transformation takes advantage of conjugate reduction of an α,β-unsaturated aldehyde by Stryker's reagent[12] and its subsequent protonation through an internal alcohol.[13] The synthesis of the required enal **4** is achieved via vinylogous Mukaiyama aldol[14] reaction (VMAR) of the chiral vinylketene silyl *N,O*-acetal **5** an aldehyde **6** affording the *anti* aldol product **7** (Figure 4).

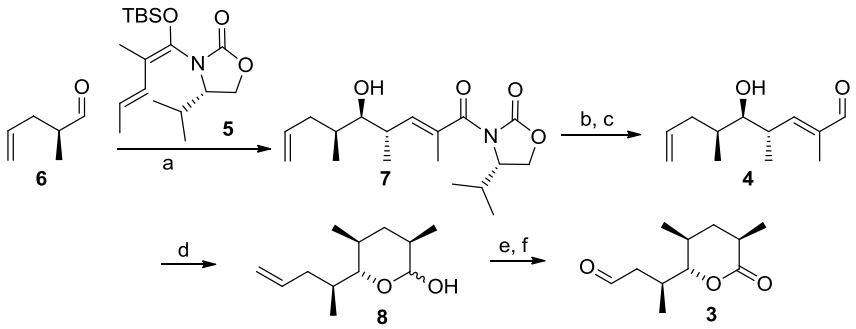


*Figure 4: Synthesis of aldehyde **3** and postulated mechanism of the stereoselective intramolecular protonation. a)$TiCl_4$, CH_2Cl_2, –78 °C to –30 °C, 18 h, 61% (two steps), d.r. 17:1; b) $LiBH_4$, THF, MeOH, 0 °C, 3 h, 90%; c) MnO_2, CH_2Cl_2, RT, 2 h, quant.; d) Stryker's reagent, benzene, RT, 15 h, 65%, d.r. 20:1; e) TPAP, NMO, MS 4 Å, CH_2Cl_2, RT, 14 h, 90%; f) O_3, CH_2Cl_2, –78 °C, 10 min, then PPh_3, RT, 1 h, quant. Stryker's reagent= $[\{(PPh_3)CuH\}_6]$, TPAP[15] = Tetrapropyl-ammonium perruthenate, NMO= N-Methylmorpholine-N-oxide, MS= molecular sieve.*

The western fragment **2** could be synthesized starting from Boc-protected amino alcohol **9** (Figure 5). Swern[16] oxidation established the aldehyde functionality required for the following sequence of two olefination reactions. The first olefination yielded aldehyde **10** which was used in the next transformation to generate ester **11**. The aldehydes were used immediately in the next steps to avoid racemization, yielding ester **11** in 80%*ee*. Liberating the amine with TFA/CH_2Cl_2 at 0 °C and peptide

coupling with acid **12** (generated from 2-oxovaleric acid in two steps[17]) gave access to compound **13**. The synthesis continued by reducing the ester functionality of **13** to the alcohol which was submitted to an Appel reaction to provide corresponding bromide.[18] Reaction with tributylphosphine provided phosphoniumbromide **2** and completed the synthesis of this key building block.

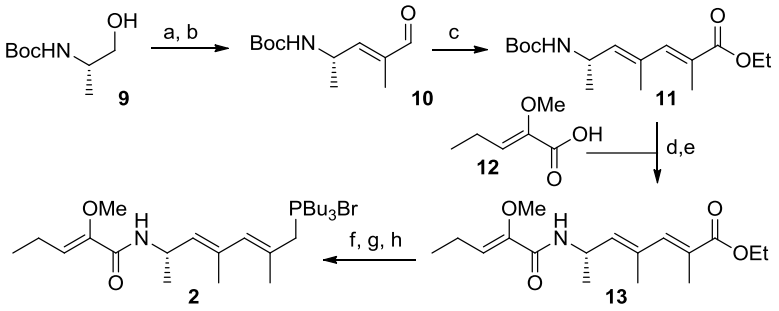


*Figure 5: Synthesis of fragment **2**. Reagents and conditions: a) $(COCl)_2$, DMSO, Et_3N, CH_2Cl_2, –78 °C to 0 °C, 2 h; b) $Ph_3P{=}C(CH_3)CHO$, toluene, 90 °C, 14 h, 65% (two steps); c) $Ph_3P{=}C(CH_3)CO_2Et$, CH_2Cl_2, RT, 16 h, 85%, 80%ee[a]; d) TFA, CH_2Cl_2, 0 °C, 30 min; e) **12**, EDC, HOBt, CH_2Cl_2, 0 °C, RT, 14 h, 82% (two steps); f) DiBAlH, CH_2Cl_2, –78 °C, 2 h, 87%; g) CBr_4, PPh_3, CH_2Cl_2, RT, 5 min, 76%; h) PBu_3, MeCN, RT, 3 h, quant. TFA= Triflouroacetic acid, EDC= 1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride, HOBt= 1-Hydroxybenzotriazole, DiBAlH= Diisobutylaluminiumhydride. [a] Determined by chiral GC.*

With both building blocks in hand we were able to complete the synthesis of pellasoren A (**1a**) by connecting both segments through an *E*-selective Wittig olefination between C8 and C9. Comparison of the ^{1}H-NMR data of synthetic and authentic pellasoren confirmed the relative configuration of pellasoren as shown in structure **1a** (see Supporting Information). Additionally, the two chromophores joined by carbon C14, should provide the desired CD data to distinguish between both enantiomers. In order to compare both isomers, additionally the unnatural diastereomer (14*R*)-pellasoren was synthesized following the same strategy as described for **1a**. Gratifyingly, comparison of the NMR and CD spectra of pellasoren (**1a**), (14*R*)-pellasoren and the authentic sample validated the absolute and relative configuration of this natural product as shown for **1a** (Figure 1).

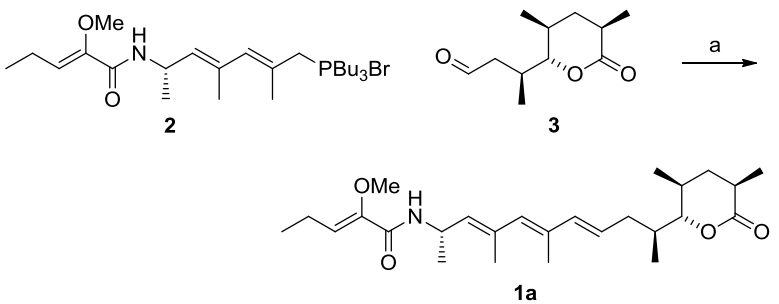


*Figure 6: Completion of pellasoren (**1a**). Reagents and conditions: a) KOtBu, THF, –78 °C to RT, 15 h, 46%.*

Validation of the pellasoren configuration by total synthesis confirms the above-described deviations from bioinformatic analysis, and thus underpins the current limits of ER sequence-based

stereochemistry prediction. When seeking to improve bioinformatic tools, the pellasoren ER domains may provide a useful starting point for future biochemical and structural studies, as they show 92.4 % identity spanning 315 aa residues where Arg43 from PelF-ER is substituted to Leu43 in PelD-ER. We hypothesize here that Arg43 could be another important amino acid to govern the stereochemistry of the reduced extender unit.

The genes involved in biosynthesis of the rarely observed glycolate extender unit are located upstream of *pel*A-F. Specifically, we propose that PelG to PelK generate the unusual "glycolate" extender unit believed to be incorporated by module 2, because these proteins show high similarity to proteins responsible for synthesis of glycolate precursors in the soraphen-, ansamitocin- and FK-520 pathways (Table 1). Unfortunately and despite serious efforts, our attempts to prove the gene cluster analysis based hypothesis by incorporation studies using isotopically labelled precursors and employing MS/MS analysis were not conclusive due to heavy metabolism of fed compounds (Supporting Figure 7).[7] However, the above mentioned similarity to known glycolate pathways strongly supports the biosynthesis as shown in Figure 2.

Table 1: Comparison of proteins involved in biosynthesis of glycolate extender units.

protein	aa	proposed function of the homologous protein	identity/similarity to	
			soraphen biosynthetic genes	FK-520 biosynthetic genes
		Acyl transferase (28-355)	57/65 % (SorC - AT part)	18/26 % (FkbH)
PelG	794	Ppant attachment site (427-480)	22/28 % (SorC - ACP part)	21/29 % (FkbJ)
		FkbM methyltransferase (589-744)	39/48 % (SorC - OMT part)	8/15 % (FkbG)
PelH	283	Hydroxyacyl-CoA dehydrogenase	74/82 % (SorD)	36/51 % (FkbK)
PelI	93	Acyl carrier protein	57/67 % (SorX)	27/44 % (FkbJ)
PelJ	384	Acyl-CoA dehydrogenase	49/60 % (SorE)	33/48 % (FkbI)
PelK	376	HAD superfamily phosphatase	not present	41/57 % (FkbH)

It is assumed that 1,3-bisphosphoglycerate is loaded to an ACP and according to soraphen biosynthesis, PelG-ACP binds 1,3-phosphoglycerate loaded by the PelG-AT domain. Methylation could then be catalysed by the internal SAM-dependent PelG-MT domain. These functions are usually facilitated by discrete proteins rather than a multifunctional protein. Moreover, the presence of the *pelK* gene is fairly unexpected. PelK shows homology to FkbH, a protein that is assumed to load 1,3-bisphosphoglycerate to FkbJ in FK-520 pathway.[5] Since PelG-AT is supposed to catalyse this loading step during pellasoren biosynthesis, PelK seems to be redundant. The *pel* gene cluster actually represents the first finding with this specific gene arrangement. However, when speculating about a hypothetical function for PelK in pellasoren biosynthesis it has to be pointed out that the precise mechanism for loading of phosphoglycerate has to date not been experimentally proven for any of the

known biosynthetic pathways which employ glycolate extender units. Thus, the presence of two proteins potentially able to catalyse the same loading reaction in the pellasoren pathway raises more questions concerning the biosynthesis of this extender unit. The involvement of *pelK* in the latter using targeted inactivation is not straight-forward due to the *pel* operon organization, but will be the subject of future studies.

Pellasoren A is cytotoxic against human colon cancer cells of the HCT-116 cell line with an IC50 of 155 nM. Interestingly, pellasoren B is around one magnitude less active against HCT-116 with an IC50 of 2.35 µM, emphasizing the importance of the linear nature of an all-(*E*)-configuration. Furthermore, both pellasorens show a strong effect on lysosomes. Upon incubation with osteosarcoma cells (U-2 OS) using an acridine orange assay, the pH of the lysosomes changed from acidic to neutral values (Supporting Figure 10). This is related to an apoptotic mechanism that remains elusive at the current state of research.[19]

This report features a new cytotoxic compound from *Sorangium cellulosum*. We present the first total synthesis of pellasoren and thereby validated the molecule's stereochemistry. A Wittig reaction in the last step of the synthesis was used to install the sensitive polyene system. The key steps include a vinylogous Mukaiyama aldol reaction of a chiral vinylketene silyl *N,O*-acetal and a stereoselective intramolecular protonation, which was shown to be an efficient and practical protocol in natural products total syntheses for the first time. Identification of the biosynthetic machinery nicely complements this work by enabling bioinformatics and chemical analyses including an unusual pathway to the glycolate extender unit.

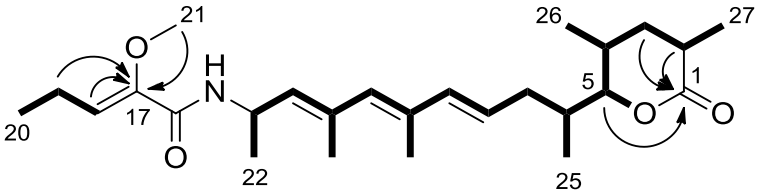
2.3 Supporting Information

2.3.1 Structure

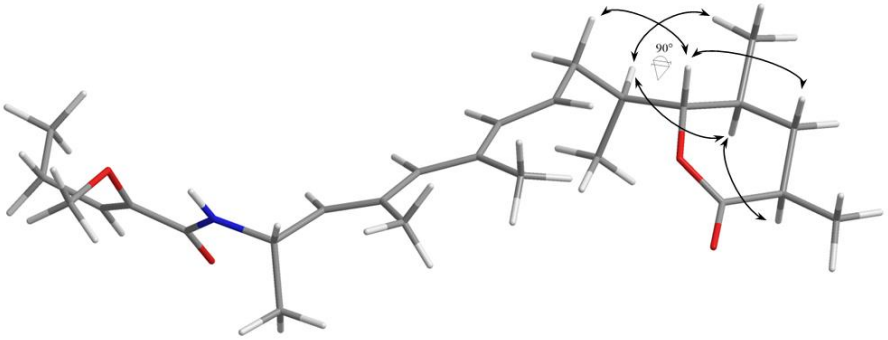
Pellasoren A and B are isobaric compounds having the same constitution and a sum formula of $C_{26}H_{41}NO_4$. The corresponding $[M+H]^+$ ion is easily detectable by ESI-MS analysis at 432.31007 *m/z* ($\Delta m/z$ = -1.7 ppm). The $[M+H]^+$ ion fragments partially during the ESI process which results in the major fragment ion of 303.23008 m/z. The structure of pellasoren A was elucidated by NMR spectroscopy and high resolution MS/MS^n analysis. The derivative pellasoren B differs by the ^{1}H-NMR shifts of H-8, H-9, H-11, and H-13 while coupling constants of H-9 and H-13 remained constant, indicating some changes in the double bond system. Same exact masses and identical fragmentation behavior of both compounds proved furthermore the conclusion of Z/E isomerism. Prolonged storage (2 weeks at room temperature) of pure pellasoren A or B samples leads to conversion between both constitutional isomers to end up with a mixture in both cases. A photochemical mechanism is the most reasonable explanation for this effect.

Supporting Table 1: *NMR spectroscopic data of pellasoren A and B. NMR spectra for structure elucidation were recorded in MeOH-d_4 at 700 MHz conducting an Ascend 700 using a cryogenically cooled triple resonance probe (Bruker Biospin, Rheinstetten, Germany).*

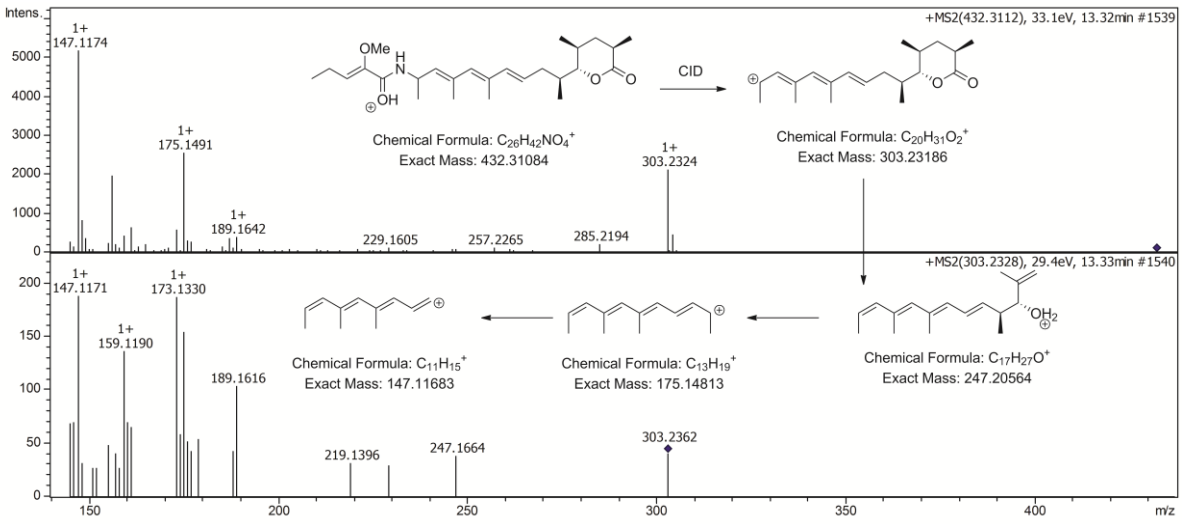
Position	Pellasoren A		Pellasoren B	
	δ_H	δ_C	δ_H	δ_C
1	-	177.6 (Cquat)	-	n.a.
2	2.54 m	37.4 (CH)	2.53 m	37.2 (CH)
3a	1.39 m	38.6 (CH2)	1.38 m	38.1 (CH2)
3b	1.91 m	38.6 (CH2)	1.94 m	38.1 (CH2)
4	1.98 m	31.6 (CH)	1.94 m	31.7 (CH)
5	4.02 dd (10.2, 1.4)	90.3 (CH)	4.01 dd (10.5, 1.4)	90.0 (CH)
6	1.92 m	36.1 (CH2)	1.94 m	35.9 (CH2)
7a	2.21 m	38.4 (CH2)	2.19 m	38.3 (CH2)
7b	2.30 m	38.4 (CH2)	2.28 m	38.3 (CH2)
8	5.67 dt (15.6, 7.3)	128.5 (CH)	5.74 dt (15.3, 7.4)	130.1 (CH)
9	6.14 d (15.4)	138.7 (CH)	6.6 d (15.5)	131.8 (CH)
10	-	135.4 (Cquat)	-	n.a.
11	5.81 s	134.1 (CH)	5.74 s	132.2 (CH)
12	-	135.3 (Cquat)	-	n.a.
13	5.34 d (8.7)	132.7 (CH)	5.28 d (8.6)	132.7 (CH)
14	-	150.2 (Cquat)	-	n.a.
15	4.86	44.8 (CH)	4.59	44.6 (CH)
16	-	164.8 (Cquat)	-	n.a.
18	6.04 t (7.6)	126.1 (CH)	6.05 t (7.7)	125.8 (CH)
19	2.24 m	19.8 (CH2)	2.24 m	19.5 (CH2)
20	1.05 t (7.6)	13.9 (CH3)	1.05 t (7.6)	13.8 (CH3)
21	3.59 s	61.0 (CH3)	3.60 s	60.8 (CH3)
22	1.26 d (6.7)	21.5 (CH3)	1.28 d (6.9)	21.2 (CH3)
23	1.83 s	17.3 (CH3)	1.79 s	17.3 (CH3)
24	1.90 s	13.9 (CH3)	1.86 s	21.2 (CH3)
25	0.91 d (6.6)	12.6 (CH3)	0.91 d (6.7)	12.2 (CH3)
26	0.97 d (6.6)	17.4 (CH3)	0.97 d (6.5)	17.3 (CH3)
27	1.24 d (7.3)	17.6 (CH3)	1.23 (6.9)	17.1 (CH3)



Supporting Figure 1: *HMBC (arrows) and 1H,1H-COSY (bold) correlations for pellasoren A.*



Supporting Figure 2: *NOE correlations (arrows) as identified by ROESY. Model generated with Chem3D Pro 12.0 running MM2 calculation to minimize energy. The bond angle H(5)-C(5)-C(6)-H(6) was fixed to 90° according to the very small coupling constant of 1.4 Hz between these two protons.*



Supporting Figure 3*: Fragmentation pattern of pellasoren A with some of the fragment ions assigned. Collision induced dissociation of the [M+H]$^+$ ion (upper part) and the main fragment 303.2301 m/z (lower part). The main fragment 303.2301 m/z is readily formed in the ESI process. Thus both fragmentations are in MS² stage for better S/N values in the resulting spectra.*

2.3.2 Analysis of the gene cluster in So ce38

The pellasoren gene cluster sequence was deposited in the EMBL database with the accession no HE616533. The region of interest as well as 20 kb flanking region on both sides was searched for open reading frames (ORFs) using Glimmer 3.0 and subsequently subjected to an automatic annotation by the CLUSEAN software tool[1;2]. The *pel* locus is 57,502 bp in size and covers eight modules within the genes

*pel*A to *pel*K with a GC content of 72.1 %. An ABC transporter gene is located at the very beginning of the cluster.

Supporting Table 2: *Genes involved in pellasoren biosynthesis as annotated in the cluster of the Sorangium cellulosum So ce38 genome.*

Gene	Start	End	Length	Function
*pel*T	1478	3736	2259	ABC transporter
*pel*G	4808	7192	2385	AT-ACP-MT
*pel*H	7189	8040	852	3-OH-acyl-CoA-dehydrogenase
*pel*I	8044	8325	282	ACP
*pel*J	8375	9529	1155	Acyl-CoA-dehydrogenase
pelK	9581	10711	1131	AT functionality
*pel*A	10735	22728	11994	ACP-KS-AT-AT-DH-KR-ACP-C-A-KR-PCP
*pel*B	22730	28303	5574	KS-AT-DH-KR-ACP
*pel*C	28317	33929	5613	KS-AT-DH-KR-ACP
*pel*D	33926	45976	12051	KS-AT-DH-KR-ACP-KS-AT-DH-ER-KR-ACP
*pel*E	45973	51621	5649	KS-AT-DH-KR-ACP
*pel*F	51618	58979	7362	KS-AT-DH-ER-KR-ACP

Supporting Table 3*: Proteins involved in pellasoren biosynthesis in the strain Sorangium cellulosum So ce38.*

NRPS/PKS part of the S. cellulosum so ce38 pellasoren biosynthetic machinery		
Protein	aa	Protein domains and their position in the sequence[1]
PelA	3997	NRPS/PKS: ACP' (9-74), KS (97-521), ATL (618-911), AT (1059-1354), DH (1421-1591), KR' (1954-2131), ACP'' (2233-2299), C (2393-2831), A (2836-3352), KR'' (3607-3784)
PelB	1857	PKS: KS (34-451), AT (556-851), DH (921-1091), KR (1454-1631), ACP (1734-1799)
PelC	1870	PKS: KS (36-454), AT (562-857), DH (925-1096), KR (1469-1646), ACP (1749-1814)
PelD	4016	PKS: KS' (38-456), AT' (563-858), DH' (920-1089), KR' (1436-1613), ACP' (1716-1782), KS'' (1809-2183), AT'' (2339-2637), DH'' (2705-2875), ER (3256-3571), KR'' (3596-3772), ACP'' (3875-3940)
PelE	1882	PKS: KS (25-443), AT (550-845), DH (914-1082), KR (1470-1647), ACP (1753-1818)
PelF	2453	PKS: KS (39-459), AT (566-861), DH (930-1100), ER (1477-1791), KR (1817-1993), ACP (2095-2161), TE (2208-2439)

Non-NRPS/PKS proteins of the So ce38 pellasoren biosynthetic machinery (putative conserved domains)				
Protein	aa	Proposed function of the homologous protein	Identity/similarity to soraphen biosynthesis	Identity/similarity to FK-520 biosynthesis
PelG	794	Acyl transferase (28-355)	57/65 % (SorC - AT part)	18/26 % (FkbH)
		Phosphopantetheine attachment site (427-480)	22/28 % (SorC - ACP part)	21/29 % (FkbJ)
		FkbM methyltransferase family (589-744)	39/48 % (SorC - OMT part)	22/32 % (FkbM)
PelH	283	Hydroxyacyl-CoA dehydrogenase	74/82 % (SorD)	36/51 % (FkbK)

PelI	93	Acyl carrier protein	57/67 % (SorX)	27/44 % (FkbJ)
PelJ	384	Acyl-CoA dehydrogenase	49/60 % (SorE)	33/48 % (FkbI)
PelK	376	HAD superfamily phosphatase	not present in soraphen cluster	41/57 % (FkbH)

[1] A: adenylation domain, AT: acyltransferase, C: condensation domain, DH: dehydratase, ER: enoylreductase, KR: ketoreductase, KS: ketosynthase, TE: thioesterase.

2.3.2.1 DH domains

According to the structur of pellasoren we can conclude that PelE-DH is an inactive domain. Within the consensus sequence of PelE-DH we observed two substitutions compared to the other DH domains, more precisely Arg38 and an Asp42 are substituted by alanine (Supp. Table 4). This replacement leads obviously to a change in local charge distribution within the consensus region and thus to the inactivation of the domain.

Supporting Table 4: *Active site analysis of the dehydratase domains (DH) involved in pellasoren biosynthesis.*

Consensus	L	x	x	H	x	x	x	G	x	x	x	x	P
PelA-DH	L	S	D	H	R	V	Q	G	E	V	V	F	P
PelB-DH	L	A	D	H	R	V	Q	G	E	V	V	F	P
PelC-DH	L	A	D	H	R	V	L	G	E	V	V	F	P
PelD-DH$_1$	L	A	D	H	R	V	Q	G	E	M	V	F	P
PelD-DH$_2$	L	S	D	H	R	V	Q	G	E	V	V	F	P
PelE-DH*	L	K	D	H	**A**	V	Q	G	**A**	V	L	F	P
PelF-DH	L	A	D	H	R	V	F	G	E	V	V	M	P

2.3.2.2 AT domains

Substrate specificity of acyl transferase (AT) domains was identified by conserved site analysis by aligning the AT domains to a reference AT from *E.coli* FAS, 1MLA (PDB 1MLA, UniProtKB P0AAI9). The main discriminant between loading of a methylmalonyl-CoA (mm-CoA) or malonyl-CoA (m-CoA) is the amino acid position 200 in 1MLA, whereas Ser200 supports mm-CoA and Phe200 supports m-CoA. The AT specificity is correct for the standard AT domains while the loading domain PelA-AT$_1$ and PelA-AT$_2$ (methoxymalonyl) were wrongly assigned, just as expected. The starter unit in pellasoren biosynthesis is a propionate although genetic analysis of PelA-AT$_1$ predicts an arginine at position 117 of the AT domain (Supp. Table 5). Yadav et.al. draw the conclusion that Arg117 is a highly conserved amino acid involved in the discrimination between mono- and dicarboxylic acid substrates.[3] This circumstance may lead to the assumption that PelA-AT$_1$ loads a methylmalonyl-CoA to the ACP domain which is then decarboxylated by the first KS, PelA-KS$_1$, resulting in a propionate bound to the ACP. On the contrary, such a biosynthetic pathway would need another KS domain within module PelA catalyzing the claisen

condensation instead. This is not the case; hence, the selectivity for monocarboxylic extender/starter units of AT domains is not solely governed by the absence of an Arg117.

Supporting Table 5: *Active site analysis of the acyl transferase domains (AT) according to Yadav et.al.*[3]

Domain	observed	predicted	11	63	90	91	92	93	94	117	200	201	231	250	255	15	58	59	60	61	62	70	72	197	198	199
PelA-AT_1	propionate	-	Q	Q	G	Q	S	M	G	R	S	H	T	N	V	W	D	T	A	R	I	Q	A	D	Y	A
PelA-AT_2	MeO-M	MM	Q	Q	G	H	S	M	G	R	S	H	T	N	V	W	E	I	D	V	A	E	A	D	V	A
PelB-AT	MM	MM	Q	Q	G	H	S	M	G	R	S	H	T	N	V	W	A	I	D	V	V	E	A	D	V	A
PelC-AT	MM	MM	Q	Q	G	H	S	M	G	R	S	H	T	N	V	W	E	I	D	V	V	E	A	D	V	A
PelD-AT_1	M	M	Q	Q	G	H	S	V	G	R	F	H	T	N	V	W	E	T	I	V	A	E	A	N	Y	A
PelD-AT_2	MM	MM	Q	Q	G	H	S	M	G	R	S	H	T	N	V	W	E	I	D	V	V	E	G	D	V	A
PelE-AT	MM	MM	Q	Q	G	H	S	M	G	R	S	H	T	N	V	W	E	I	D	V	V	E	G	D	V	A
PelF-AT	MM	MM	Q	Q	G	H	S	M	G	R	S	H	T	N	V	W	E	I	D	V	V	E	G	D	V	A

2.3.2.3 KR domains

The KR domain of module 7, PelE-KR, is responsible for the stereochemistry at C4 and C5. An alignment of KR domains allows no obvious discrimination between A- and B-type domains (Supp. Table 6).[4] The conformation as generated during biosynthesis is 2*R*,3*R* and thereby indicates a B_1-type domain (Figure 2a, module 7). However, the conserved LDD motif of B-domains is substituted by an FDN motif in PelE-KR[5]. This finding shows that the three amino acid motif is again just a rule of thumb when discriminating A- and B-type KR domains.

Supporting Table 6: *(a) Active site analysis of the ketoreductase domains (KR) involved in pellasoren biosynthesis. (b) Discrimination between A- and B-type KR domains according to Caffrey et.al. (2003).*[5] *The LDD sequence in motif I is an indicator for B-type KR domains while Pro144 and Asn148 are conserved in motif II.*[4;5]

active consensus	G	x	G	x	x	G	x	x	x	A
Pel-KR1	G	L	G	G	L	G	I	A	L	A
Pel-KR2	G	L	G	R	L	G	L	A	L	S
Pel-KR3	G	L	G	G	L	G	L	S	L	A
Pel-KR4	G	L	G	G	L	G	L	S	L	A
Pel-KR5	G	L	G	A	L	G	L	R	V	A
Pel-KR6	G	L	G	G	L	G	L	S	L	A
Pel-KR7	G	L	G	G	L	G	L	T	V	A
Pel-KR8	G	L	G	G	L	G	L	S	L	A

	region 88 - 103															region 134 - 149														
Pel-KR1	H	A	A	A	V	L	D	D	H	T	L	L	E	L	D	E	Y	S	S	A	L	S	L	F	G	S	P	G	Q	A	N	Y
Pel-KR2*	H	A	A	G	V	A	D	D	H	P	L	L	D	L	D	E	C	S	S	S	A	S	L	F	G	A	P	G	E	A	S	S
Pel-KR3	H	A	A	A	V	L	D	D	H	T	L	L	E	Q	S	E	Y	S	S	G	A	S	L	F	G	S	P	G	Q	G	N	Y
Pel-KR4	H	A	A	A	V	L	D	D	H	T	L	L	E	Q	S	E	Y	S	S	A	S	A	L	F	G	A	P	G	Q	G	N	Y
Pel-KR5	H	A	A	G	V	L	D	D	G	V	L	V	E	Q	S	A	F	S	S	L	S	S	V	L	G	S	G	G	Q	G	N	Y
Pel-KR6	H	A	A	A	V	L	D	D	H	T	L	L	E	Q	S	E	Y	S	S	A	A	S	L	F	G	S	P	G	Q	G	N	Y
Pel-KR7	H	T	A	V	N	F	D	N	H	P	V	L	G	L	S	E	Y	S	S	V	S	V	L	L	G	L	P	G	L	G	N	Y
Pel-KR8	H	A	A	A	V	L	D	D	H	T	L	L	E	Q	S	E	Y	S	S	A	S	A	L	F	G	S	P	G	Q	G	N	Y

2.3.2.4 ER domains

Kwan et.al. proposed the responsible amino acid motif in an enoyl reductase's active site.[6] Kwan states several key residues to be representative when trying to discriminate between 2*R* and 2*S* stereocenters. The active site was identified by aligning both ER domains to the *E.coli* quinone oxidoreductase (PDB 1QOR, UniProtKB 28304) which holds a tryptophan at position 52 that directs the conformation to 2*S*, whereas absence of Trp52 favors 2*R* conformation (Supp. Table 7). Pellasoren biosynthesis is not following this rule of thumb. Moreover, PelD-ER and PelF-ER, both with Leu52 lead to the synthesis of a 2*S* (PelD-ER) and a 2*R* (PelF-ER) stereocenter, respectively. Both ER show 92.4 % identity within 315 aa while Arg43 is substituted to Leu43 for PelD-ER within the consensus sequence. Up to now, the amino acid at position 43 was not shown to be important although this result speaks for a certain influence.

Supporting Table 7: *Active site analysis of the enoyl reductase domains (ER) by comparison to ER domains of known stereoselectivity.[6] The second to last amino acid of the consensus is putatively directing the stereochemistry of the sp³ carbon atom during reduction. Tryptophan at this position favors a 2S-configured product. The E.coli quinone oxidoreductase (PDB 1QOR, UniProtKB 28304) was used as alignment outgroup according to Kwan et.al.[6]*

Consensus	G	X	X	F	X_{43}	D	X	L	X	X	X	X	X_{52}	X
PelD-ER (2S)	G	L	N	F	**L**	D	V	L	L	A	L	G	M	**L**	P
PelF-ER (2R)	G	L	N	F	**R**	D	V	L	L	A	L	G	V	**L**	P
Nan_ER_4_2S	G	I	N	F	R	D	V	L	I	V	L	G	M	Y	P
Nig_ER_2_2S	G	V	N	F	R	D	A	L	I	V	L	G	M	Y	P
Mer_ER_6_2S	G	V	N	F	R	D	V	L	L	A	L	G	M	Y	P
Nan_ER_2_2S	G	L	N	F	R	D	A	L	I	A	L	D	M	Y	P
Mon_ER_4_2S	G	M	N	F	R	D	V	L	I	A	L	G	M	Y	P
Ole_ER_4_2S	G	V	N	F	R	D	V	L	L	A	L	G	M	Y	P
Fkb_ER_6_2S	G	L	N	F	R	D	V	L	I	A	L	G	T	Y	P
Rap_ER_7_2S	G	L	N	F	R	D	V	L	I	A	L	G	T	Y	P

Nan_ER_14_2S	G	V	N	F	R	D	V	L	I	A	L	G	M	Y	P
Fkb_ER_7_2S	G	L	N	F	R	D	V	L	I	A	L	G	T	Y	T
Tyl_ER_5_2S	G	V	N	F	R	D	A	L	I	A	L	G	M	Y	P
Hbm_ER_2_2S	G	Q	N	F	R	D	V	L	I	A	L	G	M	Y	E
Myc_ER_5_2S	G	L	I	F	R	D	T	L	I	A	L	G	V	Y	P
mon_ER_2_2S	G	V	N	F	R	D	V	L	I	A	L	G	M	Y	P
Rap_ER_3_2S	G	L	N	F	R	D	V	L	I	A	L	G	T	Y	P
Ery_ER_4_2S	G	V	N	F	R	D	V	L	L	A	L	G	M	Y	P
Pik_ER_4_2S	G	L	N	F	R	D	V	L	I	A	L	G	M	Y	P
Tmc_ER_6_2S	G	L	N	F	R	D	V	L	N	A	L	G	M	Y	P
Nig_ER_4_2S	G	M	N	F	R	D	A	L	I	A	V	G	M	Y	P
Gdm_ER_2_2S	G	Q	N	F	R	D	V	L	I	A	L	G	M	Y	E
Meg_ER_4_2S	G	V	N	F	R	D	V	L	L	A	L	G	M	Y	P
Chm_ER_5_2S	G	L	N	F	R	D	T	L	I	A	L	G	M	Y	P

2.3.2.5 Adenylation Domain Specifictiy

1) NRPSPredictor2 SVM prediction details:
8 Angstrom 34 AA code:
FWTTFDVSFFEAVTNLAGERDLYGPTEDTTYSTC

Predicted physicochemical class:
hydrophobic-aliphatic

Large clusters prediction:
gly,ala,val,leu,ile,abu,iva

Small clusters prediction:
gly,ala
Single AA prediction:
ala

2) NRPSPredictor2 Stachelhaus code prediction:
ala

3) Minowa HMM method A-domain
Substrate specificity prediction top hits:
Substrate: Score:
Ala 257.2
Leu 196.2
Val 194.4

Consensus Prediction: 'ala'

Supporting Figure 4: *Output of A-domain specificity prediction using antiSMASH.[7] The results clearly indicate specificity for alanine.*

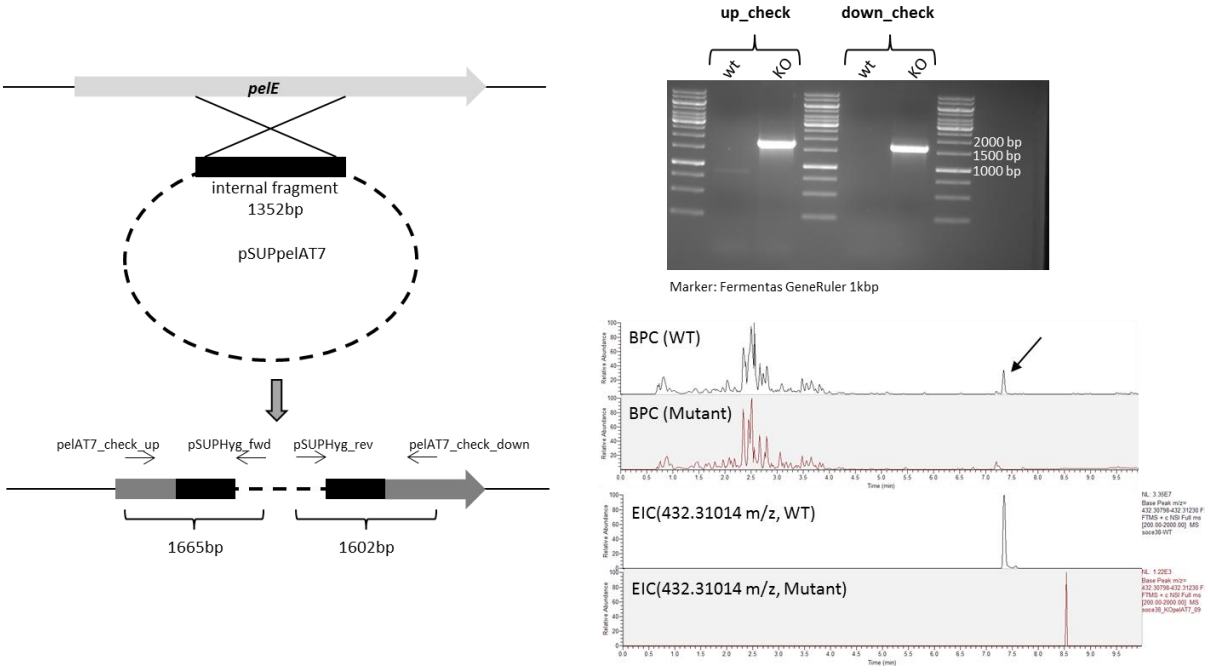
2.3.3 Knockout of pellasoren production in So ce38

Genetic modification of the So ce38 strain was accomplished according to a previously described protocol.[8] Putative single cross-over knockout mutants were grown in M-Medium (1 % papaic digest of soybean meal, 1 % maltose, 0.1 % $CaCl_2 \cdot 2\ H_2O$, 0.1 % $MgSO_4 \cdot 7H_2O$, 8 mg/l FeEDTA, 50 mM HEPES, adjusted to pH 7.2 using 10N KOH) supplemented with hygromycin (100 µg/ml) and 1% adsorber resin

(XAD-16, Rohm & Haas) for at least 5 days at 180 rpm and 30°C. Methanolic extracts of the XAD resin were analyzed by LC-MS and compared with WT extracts. For isolation of chromosomal DNA the mutant strain was grown in M-medium containing hygromycin. The integration of the plasmid at the right position in the genome was verified by PCR as shown in figure 4 in which mutant and wild-type DNA were used as template.

Supporting Table 8: *Primers used for amplification and verification of the pellasoren knock-out mutant in Sorangium cellulosum So ce38.*

oligonucleotide	sequence 5`-3`
pelAT7_fwd	GGC GAT CCC GAA GAA CCT GCA
pelAT7_rev	ACT CGC CCT CGC GGA GGT TCT
pSUPHyg_fwd	ATG TAG CAC CTG AAG TCA GCC
pSUPHyg_rev	ACG CAT ATA GCG CTA GCA GC
pelAT7_check_up	TCT TCG ATC CAG TAC CGC TC
pelAT7_check_down	ACA TCG GGT ACG TGG AGA CG



Supporting Figure 5: *Verification of KO-mutant by PCR and due to abolishment of pellasoren production as detected by LC-MS analysis of WT and mutant.*

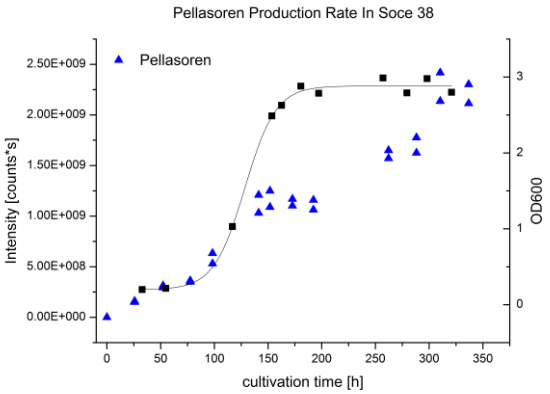
2.3.4 Isolation

Pellasoren was produced by the *Sorangium cellulosum* strain So ce38 in shaking flask cultures using M medium (10 g/l soy peptone, 10 g/l maltose, 1 g/l $CaCl_2$*$2H_2O$, 1 g/l $MgSO_4$*$7H_2O$, 11.9 g/l HEPES, 8 mg/l NaFeEDTA, adjusted to pH 7.2 using 10 N KOH). Six liters (distributed over three 5 l flasks) of M-Medium were inoculated with 600 ml of a 7-day-old So ce38 culture. After four days of incubation, 80 ml

of aqueous adsorber resin slurry (Amberlite XAD-16, 50 % (w/v)) were added to each flask. Incubation at 30° C and 150 rpm was maintained for another ten days. The cultivation broth was centrifuged at 6000 x g to separate the XAD from liquid medium and cells. 120 g XAD was washed with 2 x 500 ml of ethyl acetate followed by 2 x 500 ml of methanol/chloroform (50:50 v/v) to obtain extracts that are both equally rich in pellasoren. Extracts were combined to give 5.31 g crude extract. The dry extract was purified using silica chromatography. The column was packed with hexane and elution of pellasoren was achieved using 400 ml hexane followed by 2 l of $CHCl_3$/MeOH (95:05). The fraction of interest eluted within the early second elution step. After evaporating to dryness 366 mg of pellasoren crude extract was obtained. After dilution in 2 ml of methanol the dissolved sample was centrifuged and the supernatant applied to preparative HPLC. Preparative HPLC was performed with a Waters Autopurifier system equipped with a Waters XBridge C18 150 x 19 mm, 5 µm dp column running at 25 ml/min flow rate using (A) water + 0.1 % FA and (B) methanol + 0.1 % FA as solvent system. Separation of a 300 µl injection was done under isocratic conditions at 73 % B within 16 min. Fraction collection was triggered by a mass trigger set to 432.2 m/z, resulting in collection of two peaks. Corresponding fractions of seven consecutive runs were combined and evaporated to dryness. Pellasoren A (**1a**) eluted at 8.0 min and yielded 13.8 mg. Pellasoren B (**1b**) eluted at 10.2 min and yielded 9.0 mg.

2.3.5 HPLC-MS analysis

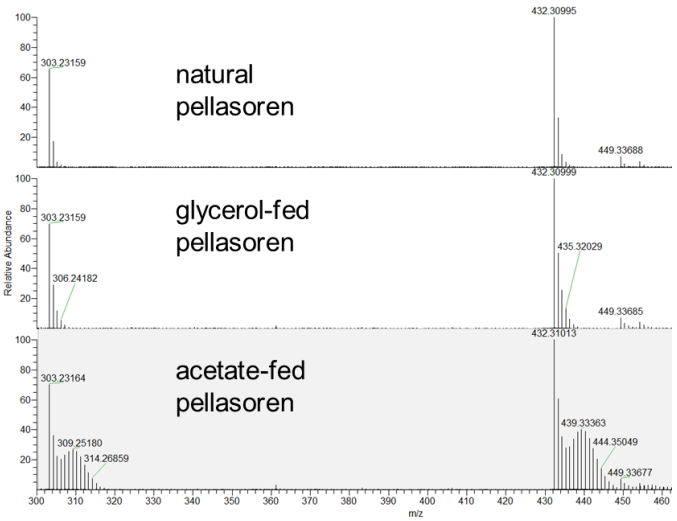
All measurements were performed on an Ultimate 3000 RSLC system (Dionex, Germering, Germany) using a Waters BEH C18, 100 x 2.1 mm, 1.7 µm dp column. Two µl of methanolic sample was injected. Separation was achieved by a linear gradient with (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 550 µl/min and 45 °C. The gradient was initiated by a 0.39 min isocratic step at 5 % B followed by an increase to 95 % B in 18 min to end up with a 1.60 min flush step at 95 % B before reequilibration using the initial conditions. Coupling the HPLC to a MS was supported by an Advion Triversa Nanomate nano-ESI system attached to an Orbitrap (ThermoFisher, Schwerte, Germany). Mass spectra were acquired in centroid mode ranging from 200 – 2000 m/z at a resolution of R = 30000 using positive electrospray ionization. Alternatively, the same HPLC setup was hyphenated to a maXis time-of-flight mass spectrometer (Bruker Daltonics, Bremen, Germany) using a mass range from 150 – 2000 m/z with positive electrospray ionization.



Supporting Figure 6: *Pellasoren production compared to cell growth. Growth state is represented by OD_{600} values. Amount of pellasoren is a relative value according to the peak height observed in HPLC-MS analysis.*

2.3.6 Feeding with labeled precursor

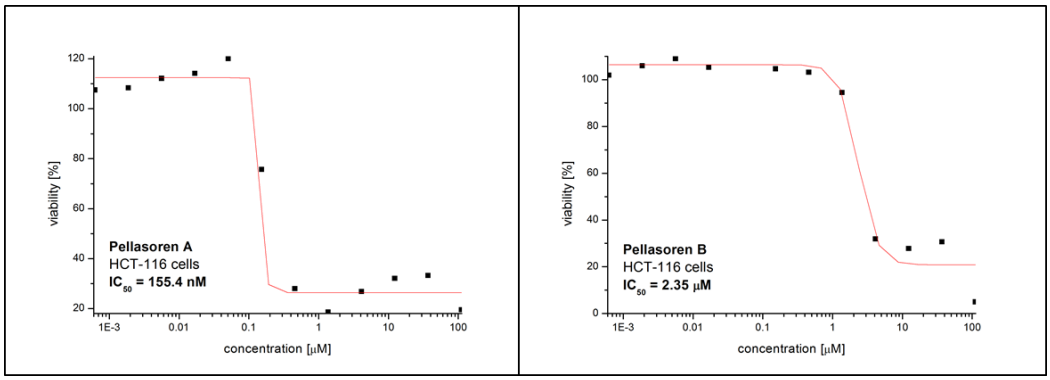
25 ml of M medium was inoculated with 2.5 ml *so ce*38 culture ($OD_{600,init}$ = 0.3) and incubated at 30 °C and 140 rpm using a 100 ml shaking flask. After two days of incubation, 0.5 ml of aqueous adsorber resin slurry (Amberlite XAD-16, 50 % (w/v)) was added. Labeled precursor was dissolved in 4 ml water, sterile filtered, and added over 4 days to end up with a final concentration of 2 mM ^{13}C-glycerol or 20 mM ^{13}C-acetate, respectively. The culture was incubated for 6 days (OD_{600} = 2.7) before extraction of the XAD resin. A reference culture was prepared the same way. Figure 7 indicates the use of labeled precursor in primary metabolism since we observe M+1 enrichments although doubly labeled precursors were fed.



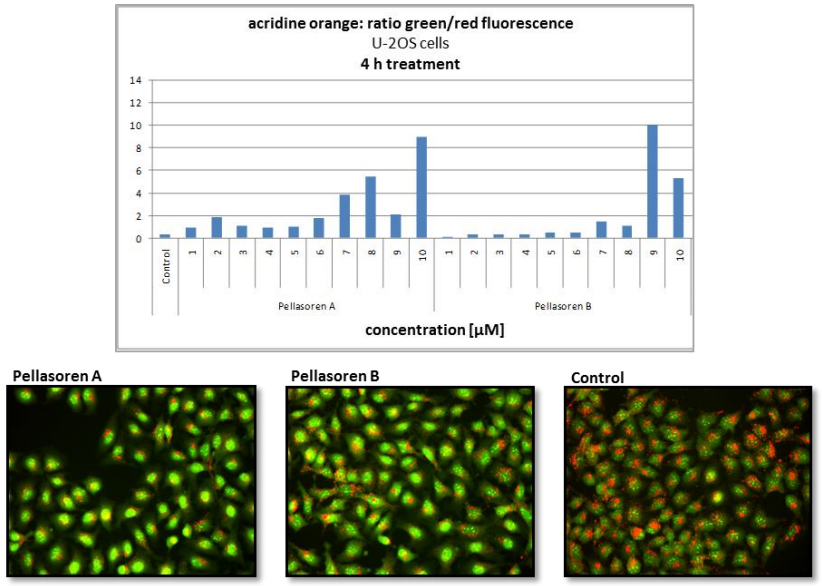
Supporting Figure 7: *Isotopic distribution of natural (upper part), glycerol-fed (middle part), and acetate-fed pellasoren (lower part). The glycerol-fed pellasoren is enriched in the M+1 isotope as well. A strict incorporation as glycerate cannot be proven from these patterns.*

2.3.7 Cytotoxicity Assay

Human HCT-116 colon carcinoma cells were seeded at 6 x 103 cells per well of 96-well plates (Corning CellBind) in 180 µl complete medium (McCoys5A /10% FBS) and directly treated with varying concentrations of pellasoren diluted in methanol. Each derivative was tested in duplicate as well as the internal methanol control. After 5 d incubation, 20 µl of 5 mg/ml 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 washed with 100 µl PBS before adding 100 µl 2-propanol/10 N HCl (250:1 v/v) in order to dissolve formazan granules. The absorbance at 570 nm was measured using a microplate reader (EL808, Bio-Tek Instruments Inc.) and cell viability was expressed as percentage relative to the respective methanol control.



Supporting Figure 8: *Cytotoxicity assay using HCT-116 cells.*

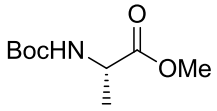


Supporting Figure 9: *Effect of pellasoren on U-2OS cells in an acridine orange assay. The acidic compartments (lysosomes) result in an orange fluorescence of the dye whereas neutral pH gives green fluorescence. Incubation with pellasoren shows an evident effect on lysosomes.*

2.3.8 Synthetic procedures:

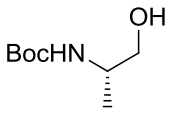
All non-aqueous reactions were carried out using flame-dried glassware under argon atmosphere. Solvents for non-aqueous reactions were dried as follows prior to use: THF was distilled from sodium/benzophenone, CH_2Cl_2 was distilled from CaH_2. All commercially available reagents and reactants were used without purification unless otherwise noted. Reactions were monitored by thin layer chromatography (TLC) using Merck Silica Gel 60 F245-plates and visualized by fluorescence under UV-light. In addition, TLC-plates were stained using a cerium sulfate/phosphomolybdic acid or $KMnO_4$ stain. Chromatographic purification of products was performed on Merck Silica Gel 60 (40-60 µm) using a forced flow of eluents. NMR-spectra were recorded on a Bruker AVS-400 or RX500 at room temperature. Chemical shifts are reported in ppm with the solvent resonance as internal standard. Data are reported as follows: s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, br. = broad signal. High–resolution mass spectra using electrospray ionization (ESI-hrMS) were recorded with a Waters Micromass LCT spectrometer with a Lock-Spray unit. Optical rotations [α] were measured on a Polarimeter 341 (Perkin Elmer) at a wavelength of 589 nm and are given in 10^{-1} deg cm^2 g^{-1}. Chiral GC experiments were carried out on a HP 5890-II device (Hewlett-Packard) with a flame ionisation detector and hydrogen as carrier gas in constant flow modus. A Hydrodex-® PM capillary column (50 m, 0.25 mm, 723370, Macherey-Nagel) was used for separation of enantiomers.

• Ester **14**



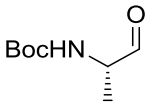
To a solution of Boc-L-alanine (5.00 g, 26.4 mmol) in DMF (40 mL), Na_2CO_3 (6.05 g, 57.1 mmol) and methyl iodide (7.00 mL, 1.12 mmol) were added. The solution was stirred for 15 h and diluted with EtOAc (20 mL). The organic layer was washed with water (30 mL) and brine (30 mL), dried over $MgSO_4$ and concentrated *in vacuo*. The residue was used directly in the next steps without further purification. Ester **14** (4.62 g, 22.7 mmol, 99%) was obtained as white solid. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 5.01 (s, 1H), 4.29-4.26 (m, 1H), 3.70 (s, 3H), 1.40 (s, 9H), 1.34 (d, J = 7.2 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 173.8, 155.1, 79.8, 52.3, 49.1, 28.3, 18.6; **HRMS (ESI):** calculated for $C_9H_{17}NO_4Na$ 226.1055, found 226.1055; **$[\alpha]^{20}_{589}$:** –3.8 (c 1.0, $CHCl_3$).

• Alcohol **9**



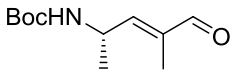
$LiAlH_4$ (1.13 g, 30.0 mmol) in THF (50 mL) was cooled to 0 °C. To this suspension, ester **14** (4.06 g, 20.0 mmol) in THF (5 mL) was added. The solution was stirred for 30 min, warmed to room temperature and stirred for additional 10 min. The mixture was quenched with water (2.80 mL), NaOH (2.80 mL, 2 mol/L) was added and the solution was stirred for 15 min. $MgSO_4$ was added, the solution was filtered and concentrated under reduced pressure. The residue was purified by flash chromatography (PE/EtOAc 3:1) and alcohol **9** (2.87 g, 16.4 mmol, 82%) was obtained as white crystals. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 4.71-4.65 (m, 1H), 3.75 (br. s, 1H), 3.62 (dd, J = 10.9, 3.8 Hz, 1H), 3.49 (dd, J = 10.9, 6.2 Hz, 1H), 2.59 (br. s, 1H), 1.43 (s, 9H), 1.13 (d, J = 6.8 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 156.3, 79.7, 67.3, 48.6, 28.4, 17.3; **HRMS (ESI):** calculated for $C_8H_{17}NO_3Na$ 198.1106, found 198.1104; **$[\alpha]^{20}_{589}$:** –9.6 (c 1.0, $CHCl_3$).

• Aldehyde **15**



$(COCl)_2$ (1.70 mL, 19.5 mmol) in CH_2Cl_2 (42 mL) was cooled to –78 °C and DMSO (2.20 g, 30.2 mmol) was added. After 15 min, alcohol **9** (2.66 g, 15.4 mmol) in CH_2Cl_2 (16.5 mL) was added and the mixture was stirred for 30 min. NEt_3 (10.5 mL, 75.7 mmol) was added dropwise and after 15 min at –78 °C, the solution was warmed to 0 °C and stirred for 30 min. The reaction mixture was diluted with water (20 mL) and layers were separated. After extraction with CH_2Cl_2 (2x 20 mL) the organic layer was washed with HCl (2x 30 mL, 2 mol/L) and sat. $NaHCO_3$ (30 mL). The solution was dried over $MgSO_4$, filtered and evaporated. The crude aldehyde was used directly in the next step. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 9.53 (s, 1H) 5.15 (s, 1H), 4.21-4.18 (m, 1H), 1.42 (s, 9H), 1.30 (d, J = 7.3 Hz, 3H), **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 199.8, 155.3, 80.0, 55.5, 28.2, 14.7; **HRMS (ESI):** calculated for $C_8H_{15}NO_3Na$ 196.0950, found 196.0951; **$[\alpha]^{20}_{589}$:** +40.8 (c 1.0, $CHCl_3$).

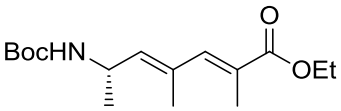
• Aldehyde **10**



To a solution of aldehyde **15** (125 mg, 0.731 mmol) in toluene (5 mL) was added 2-(triphenylphosphoranylidene)propanal (580 mg, 1.82 mmol) and the solution was stirred at 90 °C

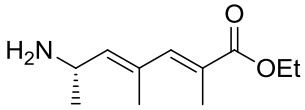
overnight. The solution was diluted with PE (5 mL), filtered and evaporated. The residue was purified by flash chromatography (PE/EtOAc 3:1) and the product **10** (100 mg, 469 mmol, 65% over 2 steps) was obtained as white solid. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 9.39 (s, 1H) 6.25 (dd, J = 8.3, 1.3 Hz, 1H), 4.66 (br. s/m, 2H), 1.81 (s, 3H), 1.42 (s, 9H), 1.27 (d, J = 6.6 Hz, 3H), **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 195.0, 155.0, 154.5, 144.8, 79.9, 28.3, 20.3, 9.2; **HRMS (ESI):** calculated for $C_{11}H_{19}NO_3Na$ 236.1263, found 236.1263; **$[\alpha]^{20}_{589}$:** +9.6 (c 1.0, $CHCl_3$).

• Ester **11**



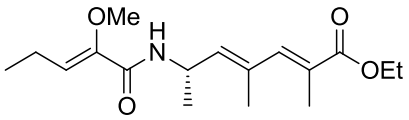
Aldehyde **10** (639 mg, 3.00 mmol) was solved in CH_2Cl_2 (12 mL) and ethyl 2-(triphenylphosphoranylidene)propanoate (2.31 g, 6.31 mmol) was added. The solution was stirred at room temperature overnight. PE (10 mL) was added, the resulting yellow solid was filtered, the solution was evaporated and the residue was purified by flash chromatography (PE/EtOAc 5:1). Ester **11** (758.3 mg, 2.55 mmol, 85%) was obtained as white solid. The enantiomeric excess was determined to be 80% *ee* by chiral GC using Hydrodex®-PM capillary column (120-200 °C, 0.5 °C/min) by comparison of the peaks at 84.1 min and 84.7 min. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 7.06 (s, 1H) 5.36 (d, J = 7.9 Hz, 1H), 4.47 (br. s/m, 2H), 4.20 (q, J = 7.1 Hz, 2H), 1.98 (s, 3H), 1.89 (s, 3H), 1.44 (s, 9H), 1.30 (t, J = 7.1 Hz, 3H), 1.21 (d, J = 6.1 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 168.8, 154.9, 141.9, 136.6, 126.9, 121.1, 79.4 60.7, 44.7, 28.4, 21.4, 16.6, 14.3, 14.0; **HRMS (ESI):** calculated for $C_{16}H_{27}NO_4Na$ 320.1838, found 320.1834; **$[\alpha]^{20}_{589}$:** -30.2 (c 1.0, $CHCl_3$).

• Amine **16**



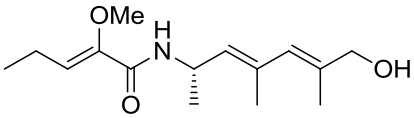
Ester **11** (200 mg, 0.673 mmol) in CH_2Cl_2 (3 mL) was cooled to 0 °C and TFA (1.5 mL) was added. After stirring at 0 °C for 30 min, the solution was reduced *in vacuo*. The free amine (**16**) was directly used without further purification. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 7.42 (br. s, 2H), 7.02 (s, 1H), 5.48 (d, J = 8.6 Hz, 1H), 4.34-4.25 (m, 1H), 4.22 (q, J = 7.1 Hz, 2H), 1.95 (s, 3H), 1.89 (s, 3H), 1.48 (d, J = 6.0 Hz, 3H), 1.32 (t, J = 7.1 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 169.2, 140.4, 138.2, 129.0, 128.2, 125.3, 61.5, 46.7, 21.5, 19.4, 16.5, 14.1, 13.7; **HRMS (ESI):** calculated for $C_{11}H_{20}NO_2Na$ 198.1494, found 198.1491; **$[\alpha]^{20}_{589}$:** –3.5 (c 1.0, $CHCl_3$).

• Ester **13**



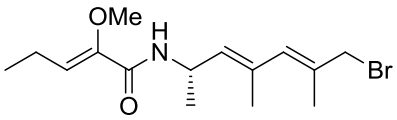
Amine **16** (394 mg, 2.00 mmol) and acid **12** (400 mg, 3.08 mmol) were solved in CH_2Cl_2 (20 mL). HOBt (400 mg, 2.10 mmol) was added and the mixture was cooled to 0 °C. EDC (500 mg, 2.61 mmol) and DIPEA (0.371 mL, 2.10 mmol) were added and the reaction mixture was stirred at room temperature overnight. HCl (10 mL, 0.1 mol/L) was added and the solution was diluted with CH_2Cl_2 (10 mL). The aqueous layer was extracted with CH_2Cl_2 (2x 20 mL), the organic layer was then washed with sat. NH_4Cl (30 mL), dried over $MgSO_4$, filtered and concentrated under reduced pressure. After flash chromatography (PE/EtOAc 5:1) compound **13** (488 mg, 1.64 mmol, 82% over 2 steps) was yielded as colourless oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 7.06 (s, 1H), 6.43 (d, *J* = 7.6 Hz, 1H), 6.19 (t, *J* = 7.7 Hz, 1H), 5.41 (d, *J* = 8.8 Hz, 1H), 4.97-4.78 (m, 1H), 4.18 (q, *J* = 7.1 Hz, 2H), 3.59 (s, 3H), 2.24-2.17 (m, 2H), 1.96 (s, 3H), 1.91 (s, 3H), 1.29 (d, *J* = 7.2 Hz, 3H), 1.27 (d, *J* = 6.9 Hz, 3H), 1.03 (t, *J* = 7.5 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 168.7, 162.5, 148.1, 141.7, 135.5, 134.0, 127.1, 125.4, 60.7, 43.2, 21.3, 19.0, 16.7, 14.2, 13.9, 13.5; **HRMS (ESI):** calculated for $C_{17}H_{27}NO_4Na$ 332.1838, found 332.1841; **$[\alpha]^{20}_{589}$:** –24.8 (c 1.0, $CHCl_3$).

• Alcohol **17**



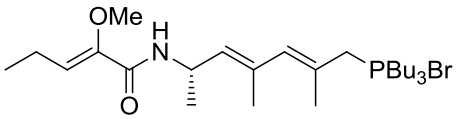
Compound **13** (140 mg, 0.497 mmol) in CH_2Cl_2 (9 mL) was cooled to *–78* °C and D*i*BAl-H (1.20 mL, 1.17 mmol, 1.0 mol/L in CH_2Cl_2) was added dropwise. After stirring for 90 min, the solution was diluted with MTBE (10 mL), warmed to room temperature and quenched with water (0.110 mL). The resulting gel was stirred and NaOH (0.40 mL, 2 mol/L) was added. After stirring for additional 10 min, $MgSO_4$ was added, the solution was filtered and reduced *in vacuo*. The residue was purified by flash chromatography (PE/ EtOAc 1:1) and alcohol **17** (109 mg, 0.432 mmol, 87%) was yielded as colourless oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 6.40 (d, *J* = 7.3 Hz, 1H), 6.21 (t, *J* = 7.7 Hz, 1H), 5.88 (s, 1H), 5.21 (d, *J* = 8.8 Hz, 1H), 4.86-4.79 (m, 1H), 4.04 (s, 2H), 3.60 (s, 3H), 2.26-2.18 (m, 2H), 1.84 (s, 3H), 1.80 (s, 3H) 1.62 (br. s, 1H), 1.27 (d, *J* = 6.6 Hz, 3H), 1.04 (t, *J* = 7.5 Hz, 3H), **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 162.4, 148.2, 135.9, 134.7, 131.0, 128.2, 125.3, 69.1, 60.7, 43.3, 21.7, 19.0, 17.4, 15.3, 13.5; **HRMS (ESI):** calculated for $C_{15}H_{25}NO_3Na$ 290.1732, found 290.1732; **$[\alpha]^{20}_{589}$:** -31.3 (c 1.0, $CHCl_3$).

• Allyl bromide **18**



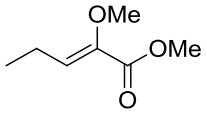
To a solution of alcohol **17** (20.0 mg, 74.5 µmol) in CH_2Cl_2 (5 mL) was added successively PPh_3 (39.0 mg, 0.149 mmol), NEt_3 (21.0 µL, 0.149 mmol) and CBr_4 (50.0 mg, 0.149 mmol). The yellow solution was stirred for 5 min, PE (5 mL) was added, the resulting solid was removed and the residue was concentrated under reduced pressure. Flash chromatography yielded bromide **18** (18.8 mg, 56.6 µmol, 76%) as yellow oil which was used directly in the next step. **HRMS (ESI):** calculated for $C_{15}H_{25}NO_2$ 330.1069, found 330.1069.

• Phosphoniumbromide **2**



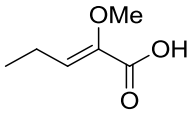
To a solution of allyl bromide **18** (15.0 mg, 45.6 µmol) in acetonitrile (1 mL), PBu_3 (56.0 µL, 0.227 mmol) was added. The mixture was stirred for 3 h, evaporated and used directly in the olefination reaction to generate pellasoren.

• Ester **19**



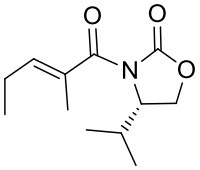
To NaH (1.03 g, 25.8 mmol) in DMF (5 mL) was added 2-oxovaleric acid (500 mg, 4.31 mmol) dropwise. The resulting foam was stirred for 90 min at room temperature. The solution was cooled to 0 °C and methyl sulfate (3.27 mL, 34.5 mmol) was added carefully over 60 min by syringe pump. The solution was stirred at room temperature for 3 h. Afterwards, the reaction was quenched with water (6 mL), the aqueous layer was extracted with diethylether (3x 10 mL). The organic layers were dried over $MgSO_4$, filtered and evaporated at room temperature. The residue was purified by flash chromatography (pentane/Et_2O 10:1) and product **19** (447 mg, 3.31 mmol, 72%) was obtained as slight yellow oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 6.24 (t, J = 7.6 Hz, 1H), 3.76 (s, 3H), 3.64 (s, 3H), 2.28-2.20 (m, 2H)., 1.02 (t, J = 7.6 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 164.3, 145.5, 130.7, 60.1, 51.7, 18.9, 13.2.

• Acid **12**



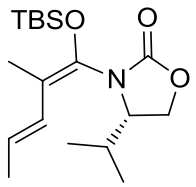
Compound **19** (100 mg, 0.69 mmol) was solved in a THF/H_2O/MeOH (2 mL, 1 mL, 0.5 mL, 4:2:1) and aq. LiOH (1 mL, 0.5 mol/L) was added. The solution was stirred for 16 h and aq. HCl (0.5 mL, 2 mol/L) was added. After extraction with Et_2O (3x 10 mL), the organic layer was dried over $MgSO_4$, filtered and concentrated under reduced pressure. Acid **12** (92.0 mg, 69.2 µmol, 100%) was obtained as colourless oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 11.34 (br. s, 1H), 6.43 (t, *J* = 7.7 Hz, 1H), 3.68 (s, 3H), 2.32–2.25 (m, 2H), 1.06 (t, *J* = 7.6 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 168.9, 145.0, 133.4, 60.3, 19.2, 13.2.

• Compound **20**



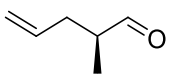
(*E*)-Methyl-2-pentenoic acid (0.36 mL, 3.09 mmol) in THF (80 mL) was cooled to –78 °C and NEt_3 (0.80 mL, 6.19 mmol) was added. After 5 min, PivCl (0.45 mL, 3.72 mmol) was added and the solution was warmed to room temperature and stirred for 60 min. LiCl (427 mg, 9.29 mmol) and (S)-4-isopropyloxazolidin-2-one (400 mg, 3.09 mmol) were added successively and the mixture was stirred for 15 h. Water (20 mL) was added, layers were separated and the aqueous layer was extracted with MTBE (3x 20 mL). The organic layer were washed with brine (2x 20 mL), dried over $MgSO_4$, filtered and evaporated. The residue was purified by flash chromatography (PE/EtOAc 6:1) and product **20** (585 mg, 2.59 mmol, 84%) was obtained as colourless oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 6.05 (t, *J* = 7.2 Hz, 1H), 4.57-4.45 (m, 1H), 4.29 (t, *J* = 8.9 Hz, 1H), 4.15 (dd, *J* = 9.0, 5.5 Hz, 1H), 2.42-2.30 (m, 1H), 2.22-2.13 (m, 2H), 1.88 (s, 3H), 1.03 (t, *J* = 7.6 Hz, 3H), 0.90 (d, *J* = 6.8 Hz, 3H), 0.88 (d, *J* = 6.4 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 171.9, 153.6, 141.1, 130.2, 63.3, 58.2, 28.2, 21.61, 17.8, 15.0, 13.4, 12.7; **HRMS (ESI):** calculated for $C_{12}H_{19}NO_3Na$ 248.1263, found 248.1260; **$[\alpha]^{20}_{589}$:** +84.2 (c 1.0, $CHCl_3$).

• Vinylketene silyl *N,O*-acetal **5**



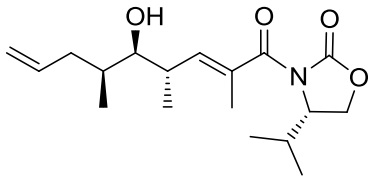
To a solution of compound **20** (275 mg, 1.22 mmol) in THF (12 mL) was added at –78 °C NaHMDS (1.22 mL, 1.83 mmol, 1.5 mol/L in THF). The solution was stirred for 90 min, TBSCl (551 mg, 3.66 mmol) was added and the reaction mixture was stirred additional 90 min. The reaction was quenched with sat. NH_4Cl (10 mL) and the aqueous layer was extracted with MTBE (3x 10 mL). The organic layers were washed with water (20 mL) and brine (20 mL), dried over $MgSO_4$, filtered and reduced *in vacuo*. The resulting residue was purified by flash chromatography (PE/ EtOAc 8:1) and product **5** (339 mg, 1.00 mmol, 82%) was yielded as white solid. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 6.22 (d, J = 15.4 Hz, 1H), 5.72-5.57 (m, 1H), 4.35-4.31 (m, 1H), 4.16-4.11 (m, 1H), 4.03-3.00 (m, 1H), 1.99-1.91 (m, 1H), 1.79 (s, 3H), 1.80-1.78 (m, 3H), 0.99 (s, 9H), 0.94 (d, J = 6.9 Hz, 6H), 0.17 (d, J = 20.0 Hz, 6H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 155.9, 134.6, 128.1, 124.3, 114.9, 64.4, 59.3, 29.3, 28.3, 26.4, 25.6, 18.7, 18.3, 16.3, 12.3, -3.6, -4.4; **HRMS (ESI):** calculated for $C_{18}H_{29}NO_4Na$ 346.1994, found 346.1992; **$[\alpha]^{20}_{589}$:** –30.2 (c 1.0, $CHCl_3$).

• Aldehyde **6**



To a solution of (*S*)-2-methylpent-4-en-1-ol (60.0 mg, 0.591 mmol) in DMSO (3 mL) was added IBX (250 mg, 0.891 mmol) and the solution was stirred for 2 h. Water (2 mL) was added, the resulting solid was filtered, the aqueous layer was extracted with Et_2O (20 mL), the organic layer was washed intensively with $NaHCO_3$ (3x 10 mL, 4 mol/L), dried over $MgSO_4$, filtered and carefully concentrated. The crude aldehyde was used directly in the next step without further purification.

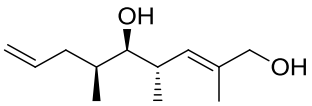
• Aldol product **7**



A solution of aldeyhde **6** (20.0 mg, 0.204 mmol) in CH_2Cl_2 (2 mL) was cooled to –78 °C and $TiCl_4$ (0.204 mL, 0.204 mmol, 1 mol/L in CH_2Cl_2) was added dropwise. Vinylketene silyl *N,O*-acetal **5** (104 mg, 0.306 mmol) was added and the solution was warmed to –30 °C. After stirring for 14 h, sat. sodium

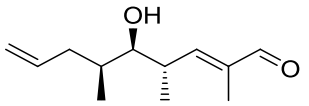
tartrate (5 mL) and sat. $NaHCO_3$ (5 mL) were added and the mixture was warmed to room temperature. The resulting slurry was stirred until the solution was nearly clear, the aqueous layer was extracted with CH_2Cl_2 (2x 10 mL), The organic phase was washed successively with water (10 mL) and brine (10 mL), dried over $MgSO_4$, filtered and evaporated. The residue was purified by flash chromatography (PE/EtOAc 5:1). Product **7** (40.3 mg, 124 µmol, 61% over 2 steps, *d.r.* 17:1) was obtained as colourless oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 5.87-5.80 (m, 1H), 5.79-5.76 (m, 1H), 5.07, 4.98 (m, 1H), 4.60-4.55 (m, 1H), 4.34 (t, J = 9.0 Hz, 1H), 4.18 (dd, J = 9.1, 2.5 Hz, 1H), 2.76-2.70 (m, 1H), 2.37-2.31 (m, 1H), 2.26-2.21 (m, 1H), 2.15-2.08 (m, 1H), 1.95 (d, J = 1.5 Hz, 3H), 1.78-1.73 (m, 1H), 0.94-0.91 (m, 12H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 171.6, 154.5, 142.3, 137.9, 131.2, 115.7, 76.6, 63.4, 58.1, 39.1, 37.4, 34.1, 28.4, 17.9, 15.6, 15.2, 13.9, 11.9; **HRMS (ESI):** calculated for $C_{18}H_{29}NO_4Na$ 346.1994, found 346.1993; **$[\alpha]^{20}_{589}$:** +20.4 (c 1.0, $CHCl_3$).

• Diol **21**



To a solution of compound **7** (80.0 mg, 0.242 mmol) in THF (20 mL) and MeOH (0.4 mL) which was cooled to 0 °C, $LiBH_4$ (0.10 mL, 0.371 mmol, 4 mol/L in THF) was added and the mixture was stirred for 3 h. Afterwards, NaOH (10 mL, 1 mol/L) was added, the mixture was stirred for 10 min at room temperature, water (10 mL) was added and the aqueous layer was extracted with CH_2Cl_2 (2x 20 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated under reduced pressure. The residue was purified by flash chromatography (PE/EtOAc 2:1) and diol **21** (42.8 mg, 0.221 mmol, 90%) was obtained as colourless oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 5.84-5.74 (m, 1H), 5.31 (d, J = 9.8 Hz, 1H), 5.07-4.99 (m, 2H), 4.03 (s, 2H), 3.26 (dd, J = 8.2, 3.3 Hz, 1H), 2.62-2.53 (m, 1H), 2.22-2.16 (m, 1H), 2.09-2.02 (m, 1H), 1.71 (s, 3H), 1.55 (bs, 2H), 0.92 (d, J = 2.3 Hz, 3H), 0.90 (d, J = 2.4 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 137.5, 136.8, 128.4, 115.9, 77.5, 68.5, 38.9, 35.9, 34.4, 17.0, 14.2, 12.3; **HRMS (ESI):** calculated for $C_{12}H_{22}O_2Na$ 221.1517, found 221.1517; **$[\alpha]^{20}_{589}$:** –7.9 (c 1.1, $CHCl_3$).

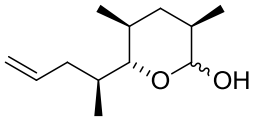
• Aldehyde **4**



To alcohol **21** (76.0 mg, 0.382 mmol) in CH_2Cl_2 (9 mL) was added MnO_2 (1.00 g, 11.5 mmol) and the suspension was stirred for 2 h. The mixture was filtered through celite® and the solution was concentrated *in vacuo*. Aldehyde **4** (74.2 mg, 0.381 mmol, 99%) was obtained as colourless oil. **^{1}H-**

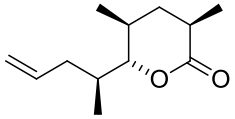
NMR (400 MHz, $CDCl_3$) δ 9.44 (s, 1H), 6.51 (d J = 9.8 Hz, 1H), 5.83-5.73 (m, 1H), 5.08-5.03 (m, 2H), 3.49-3.45 (m, 1H), 2.93-2.87 (m, 1H), 2.21-2.14 (m, 1H), 2.06-1.99 (m, 1H), 1.79 (s, 3H), 1.78-1.74 (m, 1H), 1.06 (d, J = 6.8 Hz, 3H), 0.94 (d, J = 6.8 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 195.4, 156.6, 139.5, 136.7, 116.5, 77.5, 38.5, 37.0, 35.3, 16.5, 13.0, 9.6; **HRMS (ESI):** calculated for $C_{12}H_{20}O_2Na$ 219.1361, found 219.1360; **$[\alpha]^{20}_{589}$:** +0.9 (c 1.0; $CHCl_3$)

• Lactol **8**



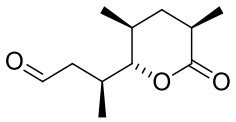
Stryker's reagent (268 mg, 0.136 mmol) in degassed benzene (7.5 mL) was stirred for 40 min under argon atmosphere. Aldehyde **7** (75.0 mg, 0.381 mmol) in degassed benzene (4 mL) was added and the solution was stirred for 14 h. The reaction was quenched by adding sat. NH_4Cl (10 mL) and the mixture was stirred for 45 min exposed to air. The resulting blue solution was extracted with CH_2Cl_2 (3x 10 mL). The organic layer was dried over $MgSO_4$, filtered and evaporated. The residue was purified by flash chromatography (PE/EtOAc 10:1). Lactol **8** (18.1 mg, 0.911 mmol, 65%) was yielded as a mixture of both isomers.

• Lactone **22**



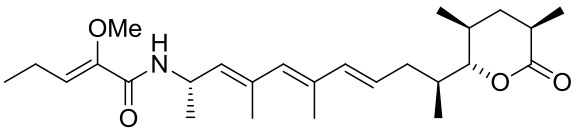
To activated molecular sieve (55 mg, 4 Å) in CH_2Cl_2 (2 mL) were added successively hemiacetal **8** (22.0 mg, 0.112 mmol) and NMO (31.5 mg, 0.269 mmol) The solution was stirred for 30 min, TPAP (2.00 mg, 5.50 µmol) was added and the solution was stirred for additional 14 h. The solution was concentrated under reduced pressure and the residue was purified by flash chromatography (PE/EtOAc 8:1). Lactone **22** (19.4 mg, 100 µmol, 90%, *d.r.* 20:1) was generated as colourless oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 5.74-5.70 (m, 1H), 5.09-5.01 (m, 2H), 3.95 (d, J = 10.2 Hz, 1H), 2.51-2.44 (m, 1H), 2.31-2.23 (m, 1H), 2.19-2.12 (m, 1H), 1.93-1.87 (m, 2H), 1.81-1.79 (m, 1H), 1.40-1.31 (m, 1H), 1.27 (d, J = 7.16 Hz, 3H), 0.95 (d, J = 6.4 Hz, 3H), 0.89 (d, J = 6.8 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 174.7, 136.8, 116.6, 88.4, 38.0, 37.6, 36.3, 34.2, 30.9, 17.2, 17.0, 12.0; **HRMS (ESI):** calculated for $C_{12}H_{20}O_2Na$ 219.1361, found 219.1359; **$[\alpha]^{20}_{589}$:** +39.8 (c 1.0; $CHCl_3$)

• Aldehyde **3**



Lactone **22** (25.0 mg, 0.127 mmol) in CH_2Cl_2 (5 mL) was cooled to –78 °C and ozone was bubbled through the solution until blue colour occured. PPh_3 (100 mg, 0.407 mmol) was added and the solution was warmed to room temperature and stirred for 60 min. The solvent was removed under reduced pressure and the residue was purified by flash chromatography (PE/EtOAc 1:1) to afford aldehyde **3** (25.7 mg, 0.127 mmol, quant.) as colourless oil. **^{1}H-NMR** (400 MHz, $CDCl_3$) δ 9.79 (s, 1H), 3.94 (dd, *J* = 10.3, 1.6 Hz, 1H), 2.79 (dd, *J* = 18.5, 6.7 Hz, 1H), 2.58 (dd, *J* = 18.5, 6.7 Hz, 1H), 2.52-2.44 (m, 2H), 1.97-1.86 (m, 2H), 1.41-1.29 (m, 1H), 1.26 (d, *J* = 7.1 Hz, 3H), 1.01 (d, *J* = 6.5 Hz, 3H), 0.93 (d, *J* = 6.8 Hz, 3H); **^{13}C-NMR** (100 MHz, $CDCl_3$) δ 201.5, 174.4, 88.6, 48.0, 37.6, 36.3, 30.9, 28.4, 17.2, 16.8, 12.2; **HRMS (ESI):** calculated for $C_{11}H_{18}O_3Na$ 221.1154, found 221.1154; **$[\alpha]^{20}_{589}$:** –5.8 (c 1.0; $CHCl_3$)

• (14*S*)-Pellasoren **1a**

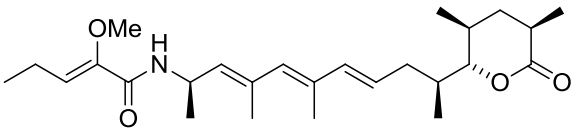


Phosphoniumbromide **2** (0.111 mmol) was solved in THF (1 mL) and cooled to –78 °C. KO*t*Bu (0.133 mL, 0.133 mmol, 1 mol/L in THF) was added dropwise, the solution was warmed to 0 °C and stirred for 45 min. Aldehyde **3** (22.0 mg, 0.111 mmol) in THF (1 mL) was added and the mixture was stirred at room temperature for 15 h. The reaction was quenched with water (2 mL), extracted with CH_2Cl_2 (2x 5 mL), the organic layer was washed with brine (10 mL), dried over $MgSO_4$, filtered and evaporated. The residue was purified by flash chromatography (PE /EtOAc 5:1) and (14*S*)-Pellasoren **1a** (21.8 mg, 50.6 µmol,46%) was obtained as a white solid. **^{1}H-NMR** (500 MHz, MeOD) δ 6.16 (d, *J* = 15.5 Hz, 1H), 6.06 (t, *J* = 7.7 Hz, 1H), 5.83 (s, 1H), 5.69 (dt, *J* = 15.1, 7.4 Hz, 1H), 5.35 (d, *J* = 8.9 Hz, 1H), 4.89-4.85 (m, 1H), 4.04 (dd, *J* = 10.5, 1.5 Hz, 1H), 3.61 (s, 3H), 2.60-2.52 (m, 1H), 2.35-2.28 (m, 1H), 2.27-2.22 (m, 2H), 2.23-2.18 (m, 2H), 2.04-1.97 (m, 1H), 1.97-1.92 (m, 2H), 1.91 (s, 3H), 1.85 (s, 3H), 1.41 (dd, *J* = 25.0, 12.4 Hz, 1H), 1.28 (d, *J* = 6.7 Hz, 3H), 1.26 (d, *J* = 7.1 Hz, 3H), 1.07 (t, *J* = 7.6 Hz, 3H), 0.99 (d, *J* = 6.5 Hz, 3H), 0.93 (d, *J* = 6.8 Hz, 3H); **^{13}C-NMR** (125 MHz, $CDCl_3$) δ 177.7, 164.8, 150.1, 138.7, 135.3, 134.1, 132.7, 128.5, 126.1, 90.3, 61.0, 44.8, 38.6, 37.4, 36.1, 32.0, 21.5, 19.8, 17.7, 17.4, 17.3, 14.3, 13.9, 12.6; **HRMS (ESI):** calculated for $C_{26}H_{41}NO_4Na$ 454.2933, found 454.2926; **$[\alpha]^{20}_{589}$:** –96.0 (c 0.105; MeOH)

• Authentic pellasoren

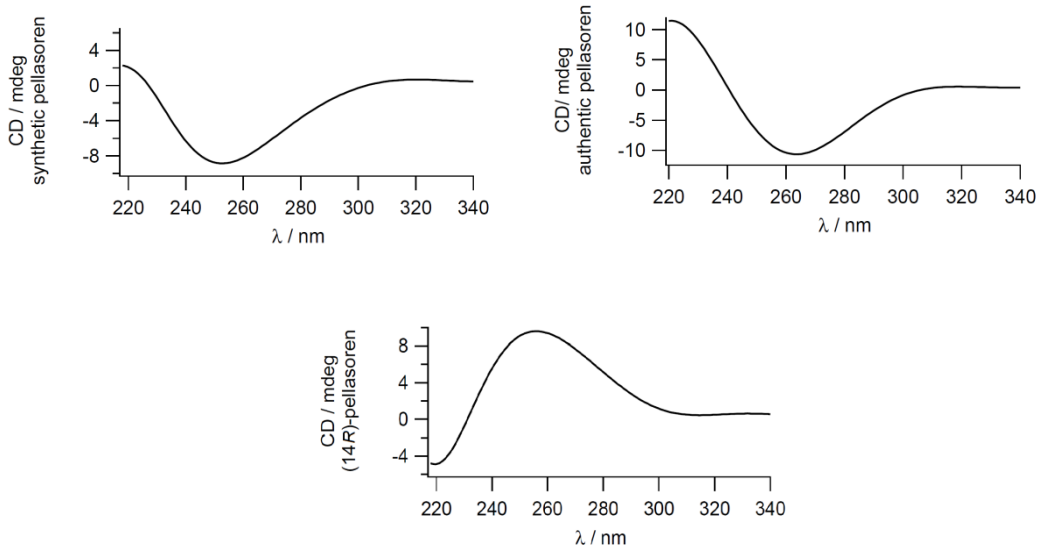
^{1}H-NMR (500 MHz, MeOD) δ 6.16 (d, *J* = 15.5 Hz, 1H), 6.06 (t, *J* = 7.7 Hz, 1H), 5.83 (s, 1H), 5.69 (dt, *J* = 15.1, 7.4 Hz, 1H), 5.35 (d, *J* = 9.1 Hz, 1H), 4.89-4.85 (m, 1H), 4.04 (dd, *J* = 10.5, 1.5 Hz, 1H), 3.61 (s, 3H), 2.59-2.52 (m, 1H), 2.35-2.28 (m, 1H), 2.27-2.22 (m, 2H), 2.23-2.18 (m, 1H), 2.04-1.97 (m, 1H), 1.97-1.92 (m, 1H), 1.91 (s, 3H), 1.85 (s, 3H), 1.41 (dd, *J* = 25.0, 12.3 Hz, 1H), 1.28 (d, *J* = 6.7 Hz, 3H), 1.26 (d, *J* = 7.1 Hz, 3H), 1.07 (t, *J* = 7.6 Hz, 3H), 0.99 (d, *J* = 6.5 Hz, 3H), 0.93 (d, *J* = 6.9 Hz, 3H); **^{13}C-NMR** (125 MHz, $CDCl_3$) δ 177.6, 164.8, 150.2, 138.7, 135.3, 134.1, 132.7, 128.5, 126.1, 90.3, 61.0, 44.8, 38.5, 37.4, 36.1, 32.0, 21.5, 19.8, 17.6, 17.4, 17.3, 14.3, 13.9, 12.6; **HRMS (ESI):** calculated for $C_{26}H_{41}NO_4Na$ 454.2933, found 454.2933; **$[\alpha]^{20}_{589}$:** –86.0 (c 0.109, MeOH)

• (14*R*)-Pellasoren



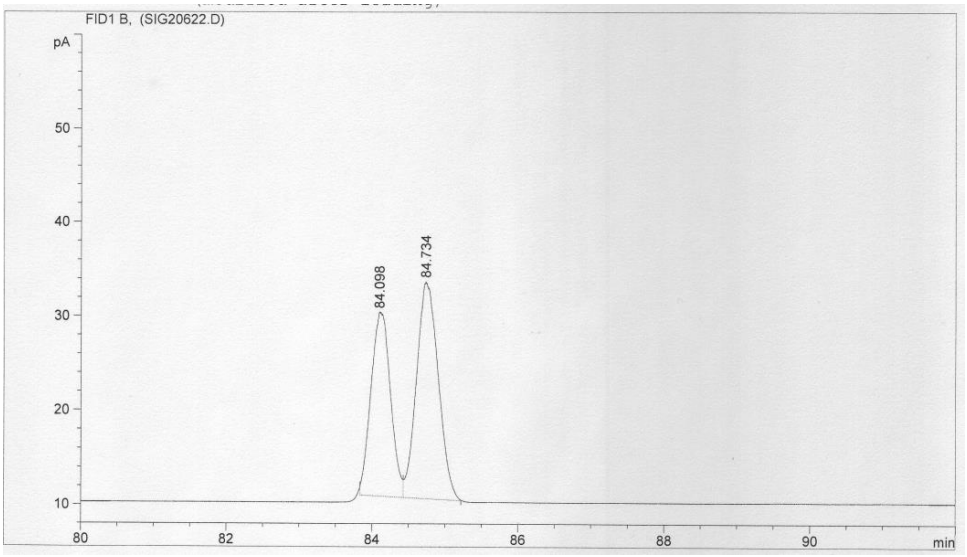
^{1}H-NMR (400 MHz, MeOD) δ 6.15 (d, *J* = 15.3 Hz, 0.5H), 6.11 (d, *J* = 15.4 Hz, 0.5H) 6.04 (t, *J* = 7.7 Hz, 1H), 5.81 (s, 1H), 5.71-5.63 (m, 1H), 5.33 (d, *J* = 8.8 Hz, 1H), 4.89-4.83 (m, 1H), 4.04 (dd, *J* = 10.4, 1.7 Hz, 0.5H), 4.02 (dd, *J* = 10.6, 1.5 Hz, 0.5H), 3.59 (s, 3H), 2.76-2.69 (m, 0.5H), 2.57-2.51 (m, 0.5H), 2.32-2.28 (m, 1H), 2.27-2.23 (m, 2H), 2.23-2.18 (m, 1H), 2.04-1.97 (m, 1H), 1.97-1.92 (m, 1H), 1.89 (s, 3H), 1.83 (s, 3H), 1.77-1.61 (m, 1H), 1.41 (dd, *J* = 24.9, 12.3 Hz, 1H), 1.28 (d, *J* = 6.7 Hz, 3H), 1.24 (d, *J* = 7.1 Hz, 1.5H), 1.16 (d, *J* = 7.1 Hz, 1.5H), 1.05 (t, *J* = 7.6 Hz, 3H), 0.99 (d, *J* = 6.8 Hz, 1.5H), 0.98 (d, *J* = 6.5 Hz, 1.5H) 0.93 (d, *J* = 6.9 Hz, 1.5H), 0.91 (d, *J* = 6.8 Hz, 1.5H); **^{13}C-NMR** (125 MHz, MeOH) δ 178.2, 177.7, 163.0, 150.3, 138.8, 138.8, 135.4, 134.2, 132.8, 132.8, 128.6, 128.5, 126.2, 90.5, 86.8, 61.1, 44.9, 38.8, 38.7, 38.5, 37.5, 36.4, 36.2, 35.8, 33.7, 32.2, 29.8, 21.6, 19.9, 17.7, 17.7, 17.5, 17.4, 17.0, 14.4, 14.0, 12.7, 12.6; **HRMS (ESI):** calculated for $C_{26}H_{41}NO_4Na$ 454.2933, found 454.2930; **$[\alpha]^{20}_{589}$:** +92.0 (c 0.103; MeOH).

2.3.9 CD Spectra

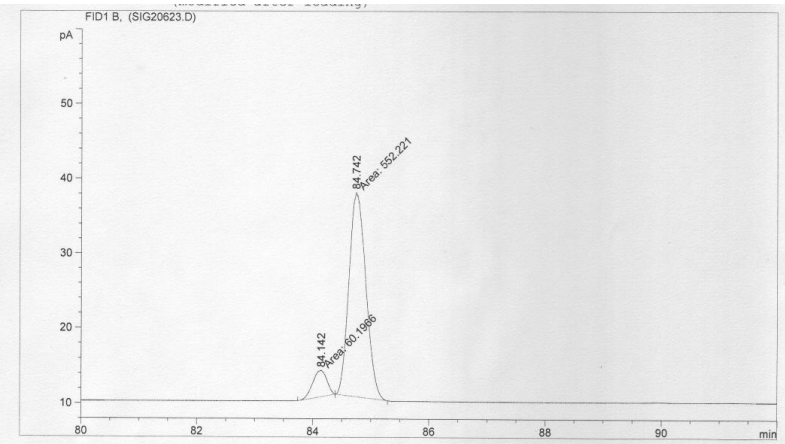


Supporting Figure 11: *CD spectra of a) synthetic pellasoren: MeOH, RT, 7.88·10^{-5} mol/L; b) authentic pellasoren: MeOH, RT, 8.6·10^{-5} mol/L; c) (14R)-pellasoren: MeOH, RT, 7.95·10^{-5} mol/L.*

2.3.10 Enantiomeric Excess of Ester 11



Supporting Figure 12: *Racemic mixture of ester **11**.*



Area Percent Report

Sorted By : Signal
Multiplier : 1.0000
Dilution : 1.0000
Use Multiplier & Dilution Factor with ISTDs

Signal 1: FID1 B,

Peak #	RetTime [min]	Type	Width [min]	Area [pA*s]	Height [pA]	Area %
1	84.142	MM	0.2837	60.19656	3.53607	9.82934
2	84.742	MM	0.3373	552.22052	27.28945	90.17066

Supporting Figure 13: *Enantiomeric excess of* ***11****.*

2.4 References

2.4.1 Main Text

[1] K. Gerth, H. Steinmetz, G. Höfle, R. Jansen, Angew. Chem. Int. Ed. 2008, 47, 600-602.
[2] S. C. Wenzel, R. Müller, Curr. Opin. Drug Discov. Devel. 2009, 12, 220-230.
[3] H. Irschik, H. Reichenbach, G. Höfle, R. Jansen, J. Antibiot. 2007, 60, 733-738.
[4] H. Reichenbach, G. Höfle, in Drug Discovery from Nature, (Eds.: S. Grabley, R. Thiericke), Springer-Verlag, Berlin, Heidelberg, 1999, p. 173.
[5] B. J. Carroll, S. J. Moss, L. Q. Bai, Y. Kato, S. Toelzer, T. W. Yu, H. G. Floss, J. Am. Chem. Soc. 2002, 124, 4176-4177.
[6] K. Wu, L. Chung, W. P. Revill, L. Katz, C. D. Reeves, Gene 2000, 251, 81-90.
[7] 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, R. Müller, H. G. Floss, J. Am. Chem. Soc. 2006, 128, 14325-14336.
[8] S. C. Wenzel, R. Müller, Nat. Prod. Rep. 2009, 26, 1385-1407.
[9] B. O. Bachmann, J. Ravel, Methods in Enzymology 2009, 458, 181-217.
[10] a) P. Caffrey, ChemBioChem 2003, 4, 654-657. b) R. Reid, M. Piagentini, E. Rodriguez, G. Ashley, N. Viswanathan, J. Carney, D. V. Santi, C. R. Hutchinson, and R. McDaniel, Biochemistry 2003, 42, 72-79.
[11] a) D. H. Kwan, Y. Sun, F. Schulz, H. Hong, B. Popovic, J. C. Sim-Stark, S. F. Haydock, P. F. Leadlay, Chem.Biol. 2008, 15, 1231-1240. b) K. Gerth, H. Steinmetz, G. Höfle, R. Jansen, Angew. Chem. Int. Ed. 2008, 47, 600-602.
[12] a) W. S. Mahony, D. M. Brestensky, J. M. Stryker, J. Am. Chem. Soc. 1988, 110, 291-293; b) B. H. Lipshutz, J. M. Serevesko, B. R. Taft, J. Am. Chem. Soc. 2004, 126, 8352-8353.
[13] A. Kena Diba, C. Noll, M. Richter, M. T. Gieseler, M. Kalesse, Angew. Chem. Int. Ed. 2010, 49, 8367-8369.
[14] a) S. Shirokawa, M. Kamiyama, T. Nakamura, M. Okada, A. Nakazaki, S. Hosokawa, S. Kobayashi, J. Am. Chem. Soc. 2004, 126, 13604-13605. b) for a general review see: G. Casiraghi, L. Battistini, C. Curti, G. Rassu, F. Zanardi, Chem. Rev., 2011, 111, 3076-3154. c) S. Shirokawa, M. Shinoyama, I. Ooi, S. Hosokawa, A. Nakazaki, S. Kobayashi, Org. Lett. 2007, 9, 849-852. d) S. Hosokawa, K. Matsushita, S. Tokimatsu, T. Toriumi, Y. Suzuki, K. Tatsuta, Tetrahedron Lett. 2010, 51, 5532-5536.
[15] S. V. Ley, J. Norman, W. P. Griffith, S. P. Marsden, Synthesis 1994, 7, 639-666.
[16] K. Omura, D. Swern, Tetrahedron 1978, 34, 1651-1660.
[17] A. Fürstner, M. M. Domostoj, B. Scheiper, J. Am. Chem. Soc. 2005, 127, 11620-11621.
[18] R. Appel, Angew. Chem. Int. Ed 1975, 87, 801-811.
[19] a) K. Kågedal, M. Zhao, I. Svensson, U. T. Brunk, Biochem. J. 2001 359, 335–343; b) A. C. Johansson, H. Appelqvist, C. Nilsson, K. Kagedal, K. Roberg, K. Ollinger, Apoptosis 2010, 15, 527-540.

2.4.2 Supporting Information

[1] A. L. Delcher, K. A. Bratke, E. C. Powers, S. L. Salzberg, Bioinformatics 2007, 23 673-679.
[2] T. Weber, C. Rausch, P. Lopez, I. Hoof, V. Gaykova, D. H. Huson, W. Wohlleben, J.Biotechnol. 2009, 140 13-17.
[3] G. Yadav, R. S. Gokhale, D. Mohanty, Journal of Molecular Biology 2003, 328 335-363.
[4] D. H. Kwan, F. Schulz, Molecules 2011, 16 6092-6115.
[5] P. Caffrey, ChemBioChem 2003, 4 654-657.
[6] D. H. Kwan, Y. Sun, F. Schulz, H. Hong, B. Popovic, J. C. Sim-Stark, S. F. Haydock, P. F. Leadlay, Chem.Biol. 2008, 15 1231-1240.
[7] M. H. Medema, K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano, R. Breitling, Nucleic Acids Res. 2011, 39 W339-W346.
[8] M. Kopp, H. Irschik, F. Gross, O. Perlova, A. Sandmann, K. Gerth, R. Müller, J.Biotechnol. 2004, 107 29-40.

Chapter 3 – Microsclerodermin

Microsclerodermins from Terrestrial Myxobacteria: An Intriguing Biosynthesis Likely Connected to a Sponge Symbiont

*Thomas Hoffmann, Stefan Müller, Suvd Nadmid, Ronald Garcia and Rolf Müller**

Journal of the American Chemical Society, **2013**, 135 (45), 16904–16911

DOI: 10.1021/ja4054509 // PubMed ID: 24124771

Published online: 14.10.2013

3 Microsclerodermin

3.1 Abstract

The microsclerodermins are unusual peptide natural products exhibiting potent antifungal activity reported from marine sponges of the genera *Microscleroderma* and *Theonella*. We here describe a variety of microbial producers of microsclerodermins and pedeins among myxobacteria along with the isolation of several new derivatives. A retro-biosynthetic approach led to the identification of microsclerodermin biosynthetic gene clusters in genomes of *Sorangium* and *Jahnella* species, allowing for the first time insights into the intriguing hybrid PKS/NRPS machinery required for microsclerodermin formation. This study reveals the biosynthesis of a "marine natural product" in a terrestrial myxobacterium where even the identical structure is available from both sources. Thus, the newly identified terrestrial producers provide access to additional chemical diversity; moreover, they are clearly more amenable to production optimization and genetic modification than the original source from the marine habitat. As sponge metagenome data strongly suggest the presence of associated myxobacteria, our findings underpin the recent notion that many previously described "sponge metabolites" might in fact originate from such microbial symbionts.

3.2 Introduction

Natural products have a longstanding tradition as leads for the development of new medicines.[1] In addition to well-established and extensively investigated plant, fungal, and bacterial producers of secondary metabolites, newer screening campaigns increasingly include organisms from less studied taxa and previously underexploited habitats such as terrestrial myxobacteria and marine sponges.[2–5] Their potential as sources of novel chemical scaffolds has been clearly demonstrated and despite the impressive structural diversity originating from these organisms, the overall picture has emerged that structural types obtained from phylogenetically distant producers usually show little overlap.[6] However, as an exception to this general notion the production of several strikingly similar compounds by unrelated species has also been reported. Some of these findings are parallel discoveries of initially sponge-derived metabolite classes from microbial sources, leading to the assumption that the respective natural products might in fact be produced by bacterial sponge symbionts.[7–9] Support for this theory comes from the identification of filamentous bacteria growing within intercellular space inside the sponge.[8,10] However, studies which unambiguously prove the production of a "sponge metabolite" by a symbiotic bacterium are exceedingly rare.[10,11] The same holds true for marine natural products of other host organisms.[12–14] This shortcoming may be attributed to difficulties with isolation and

cultivation of symbiotic microbes under laboratory conditions. Notably, the ability to independently cultivate the "real" producer of a specific secondary metabolite holds great promise, not only for sustained production but also for improving yields using both biotechnological and genetic engineering approaches. These opportunities present an invaluable advantage when further investigating a compound of interest, as the marine organism itself usually faces critical supply limitations and is poorly amenable to genetic manipulation. Moreover, access to a microbial producer facilitates the identification of biosynthetic genes underlying the formation of the metabolite of interest - a crucial prerequisite for understanding the biosynthetic machinery and a promising step toward transferring these genes into a suitable heterologous expression host.[15]

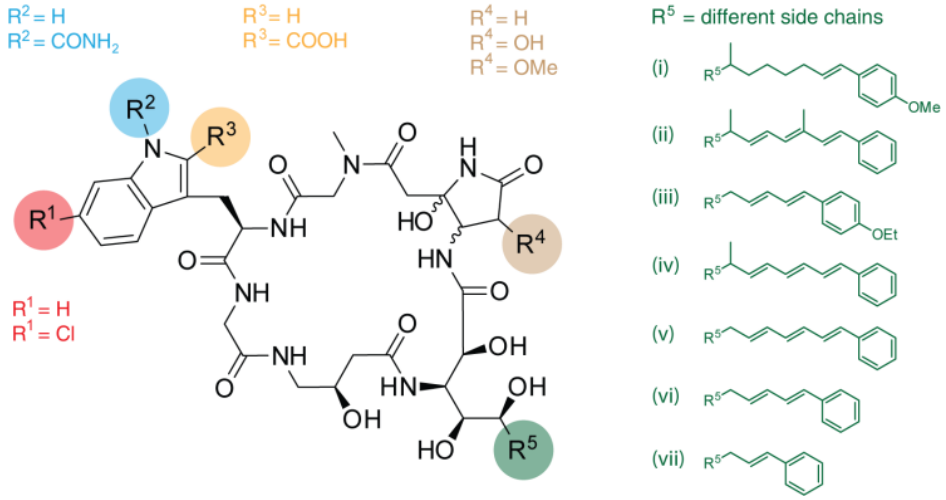


Figure 1: The microsclerodermin scaffold with an overview of the different residues identified so far. Groups R1-R4 are related to the presence of tailoring enzymes during biosynthesis whereas the side chain R5 is derived by the PKS part of the biosynthetic machinery. The pyrrolidone is reported to have R,R- or S,S-configuration, respectively.

Looking at those natural products from marine sources having apparent microbial counterparts, several cases exist where structures of myxobacterial secondary metabolites are indeed strikingly similar to previously discovered sponge-derived compounds. For example, the cyclodepsipeptide jaspamide (jasplakinolide) isolated from the marine sponge *Jaspis.*[16,17] is closely related to the structure of chondramides produced by the myxobacterium *Chondromyces crocatus* Cm c5, suggesting that the biosynthetic pathways responsible for production of these molecules should be largely similar.[18] The same holds true for renieramycin and saframycin MX1, isolated from a *Reniera* sponge and a myxobacterium of the genus *Myxococcus*.[19,20] Moreover, the macrolides salicylihalamide and apicularen were isolated from a *Halicona* sp. sponge and a *Chondromyces* species, respectively.[21,22] Very recently, bengamides were described from a *Jaspis* sponge and a cultured myxobacterium.[23,24] Adding to the list of "biosynthetic look-alikes", the structure of pedein from the terrestrial myxobacterium *Chondromyces pediculatus* Cm p3 closely resembles that of microsclerodermin,[25] which was isolated in 1994 from *Microscleroderma* sp., a lithistid sponge harvested in New Caledonia.[26] Upon their finding of pedeins in

myxobacteria, Kunze *et al.* suggested that the origin of microsclerodermins could be a bacterial sponge symbiont closely related to myxobacteria.[25] Indeed, pedein and microsclerodermin are highly similar, and both exhibit potent antifungal activity. To date several new derivatives belonging to the microsclerodermin class of peptides have been identified from various *Microscleroderma* species as well as from a *Theonella* sponge.[27–30] Nevertheless, the biosynthetic machinery behind this natural product remains so far elusive.

Table 1: Overview of Different Microsclerodermins and Pedeins and Their Origin

derivative	R^1	R^2	R^3	R^4	R^5	pyrrolidone confign	sum formula	$(M+H)^+$ [m/z]	[*a*]	[*b*]	[*c*]	[*d*]	[*e*]	ref.
A	H	H	COOH	OH	i	S, S	$C_{47}H_{62}N_8O_{16}$	995.4357	•					26,29
B	H	H	COOH	H	i	S, S	$C_{47}H_{62}N_8O_{15}$	979.4407	•					26,29
C	Cl	$CONH_2$	H	H	vii	R, R	$C_{41}H_{50}N_9O_{13}Cl$	912.3289	•	•				27
D	Cl	H	H	H	vii	R, R	$C_{40}H_{49}N_8O_{12}Cl$	869.3231	•	•	•	•		27, this study
E	H	H	COOH	H	iii	R, R	$C_{45}H_{54}N_8O_{14}$	931.3832	•					27
F + G [f]	H	H	H	H	iv	R, R	$C_{45}H_{56}N_8O_{12}$	901.4090	•					28
H + I [f]	H	H	H	H	ii	R, R	$C_{46}H_{58}N_8O_{12}$	915.4247	•					28
J	H	H	H	H	i	S, S	$C_{46}H_{60}N_8O_{12}$	917.4403	•					29
K	H	H	H	OH	i	S, S	$C_{46}H_{60}N_8O_{13}$	933.4353	•					29
L	Cl	H	H	OMe	vii	R, R	$C_{41}H_{51}N_8O_{13}Cl$	899.3337			•	•		this study
M	H	H	H	H	v	R, R	$C_{44}H_{54}N_8O_{12}$	887.3934					•	this study
Pedein A [g]	Cl	H	H	OMe	vi	R, R	$C_{43}H_{53}N_8O_{13}Cl$	925.3493			•	•		25, this study
Pedein B [g]	H	H	H	OMe	vi	R, R	$C_{43}H_{54}N_8O_{13}$	891.3883			•	•		25, this study

[a] Microscleroderma sp. (3 species) [b] Theonella sp. (1 species) [c] Chondromyces sp. (2 species) [d] Jahnella sp. (2 species) [e] Sorangium sp. (11 species) [f] The tryptophan side chain is reduced to an α-β-unsaturated amino acid. [g] Based on their same biosynthetic origin, we implicitly include pedeins when referring to the microsclerodermin family in this study.

In this study we present several terrestrial myxobacteria as alternative producers of microsclerodermins and pedeins. Our data show that *Jahnella* and *Chondromyces* species can produce the identical derivate also known from a *Microscleroderma* species. In addition, they produce new derivatives not previously reported from other sources. Access to genomic sequences for two myxobacterial producers allowed us to establish for the first time a biosynthetic model for microsclerodermin formation and also provided us with an opportunity to probe the molecular basis responsible for the structural diversity observed from microsclerodermins. Moreover, it was shown that the myxobacterial pedeins[25] originate from the same biosynthetic machinery as the microsclerodermins; hence, they belong to the same compound family. Taken together with recent metagenomic studies providing evidence that myxobacterial taxa may even exist as sponge symbionts,[31] our results underpin the assumption that a myxobacterium is the real biosynthetic source of the "marine" natural product microsclerodermin.

3.3 Experimental Section

3.3.1 Bacterial Strains and Culture Conditions

Sorangium cellulosum So ce38 was cultivated in H-medium (2 g/L soybean flour, 2 g/L glucose, 8 g/L starch, 2 g/L yeast extract, 1 g/L $CaCl_2 \cdot 2H_2O$, 1 g/L $MgSO_4 \cdot 7H_2O$, 8 mg/L Fe-EDTA, 50 mM HEPES, adjusted to pH 7.4 with 10 N KOH). Mutants of *S. cellulosum* So ce38 were cultivated in H-medium supplemented with hygromycin B (100 µg/mL) and 1 % (w/v) adsorber resin (XAD-16, Rohm & Haas) at 180 rpm and 30 °C. *Jahnella* sp. MSr9139 was cultivated in buffered yeast broth medium VY/2 (5 g/L baker`s yeast, 1 g/L $CaCl_2 \cdot 2H_2O$, 5 mM HEPES pH 7.0 with 10 N KOH) at 180 rpm and 30 °C.[32] The *Escherichia coli* strains DH10B and ET12567 harboring the plasmids pUB307 and pSUP*mscH*_KO for conjugation purposes were cultivated in Luria-Bertani (LB) medium at 37°C. Transformation of strains was performed according to the standard methods described elsewhere.[33] Antibiotics were added with the following final concentrations: chloramphenicol (25 µg/mL), kanamycin sulfate (25 µg/mL) and hygromycin B (100 µg/mL).

3.3.2 Disruption of the *msc*H *L*ocus in So ce38

Gene disruption in So ce38 using biparental mating was carried out according to a previously established protocol.[34] For construction of the plasmid pSUP*mscH*_KO a homologous fragment with the size of 2472 bp was amplified from genomic DNA using the oligonucleotides *mscH*_KO_for (GAT CCA GCG CTG GTT CCT CG) and *mscH*_KO_rev (ACG AGG CTG TCG AAG AGC G) and cloned into pCR-TOPO II-vector, resulting in the plasmid pTOPO_*mscH*_KO. The genomic segment was subsequently recovered from this plasmid using the restriction enzymes *Hind*III and *Eco*RV and further integrated into the prepared vector pSUPHyg.

3.3.3 Isolation of Microsclerodermin M from So ce38

The production medium for So ce38 was P38X medium (2 g/L peptone, 2 g/L glucose, 8 g/L starch, 4 g/L probion, 1 g/L $CaCl_2 \cdot 2H_2O$, 1 g/L $MgSO_4 \cdot 7H_2O$, 8 mg/L Fe-EDTA, 50 mM HEPES, adjusted to pH 7.5 with 10 N KOH). A 100 L fermenter with 2 % (w/v) XAD-16 adsorber resin (Rohm & Haas) was harvested after 14 days of fermentation. The cells were removed from the XAD before extraction with 3 x 3 L of methanol followed by 1 x 3 L of acetone. The combined fractions yielded 47.2 g dry weight of crude extract. Five grams of this extract was suspended in cold water, the suspension was centrifuged immediately, and the remaining pellet was dissolved in DMSO/MeOH (1:1, v/v) to give a product-enriched solution which was subjected to preparative HPLC using a Waters Autopurifier System equipped with a Waters XBridge C18, 150 x 19 mm, 5 µm d_p column operated at room temperature. The gradient started at 30 % B, increased to 50 % B in 2 min and to 51 % B in another 2 min before increasing to 95 % B in 4 min for column flushing. The combined fractions of interest were lyophilized, dissolved in

DMSO/MeOH (1:1, v/v), and forwarded to a semipreparative Dionex HPLC system (P680 pump, TCC100 thermostat, and PDA100 detector) equipped with a Phenomenex Fusion C18, 250 x 4.6 mm, 4 µm d_p column. Separation was achieved by a linear gradient using (A) H_2O and (B) ACN at a flow rate of 5 mL/min and 30 °C. The gradient started at 10 % B and increased to 30 % B in 3 min, followed by an increase to 38 % B in 15 min (0.9 % B/column volume). UV data were acquired at 316 nm. A maximum of 100 µL of the sample was manually injected before fraction collection, yielding 8.1 mg of microsclerodermin M. Microsclerodermin M: white amorphous solid, $[\alpha]_D^{20}$ - 55.7 ° (c 0.10, DMSO/MeOH 8:2).

3.3.4 Isolation of Microsclerodermins from MSr9139

The strain MSr9139 was cultivated in 3 x 1 L shaking flasks containing 500 mL of buffered VY/2 medium for 30 days. The medium was changed every 24 h by pipetting out the liquid broth. The cell pellet was harvested by centrifugation and lyophilized overnight, followed by extraction with 3 x 300 mL of methanol. The combined fractions yielded an orange-brown crude extract which was further partitioned between hexane and MeOH/H_2O 7:3 (v/v) to yield 170 mg of crude extract out of the aqueous phase. Subsequently, the extract was purified by semipreparative HPLC using an Agilent 1260 Infinity system (G1311C quaternary pump, G1330B thermostat, G1315D DAD detector and G1328C manual injector) equipped with a Phenomenex Jupiter Proteo, 250 x 10 mm, 4 µm d_p column. Separation was achieved by a linear gradient using (A) H_2O and (B) ACN at a flow rate of 2.5 mL/min and 22 °C. The gradient started at 20 % B and increased to 50 % B in 35 min (5.7 % B/column volume). UV data were acquired at 280 nm. A maximum of 100 µL of the sample was manually injected before fraction collection yielding 0.7 mg of microsclerodermin D, 0.45 mg of microsclerodermin L, and 0.85 mg of pedein A. Microsclerodermin L: white amorphous solid, $[\alpha]_D^{20}$ - 77.7 ° (c 0.12, MeOH).

3.3.5 LC-MS data acquisition

All measurements were performed on a Dionex Ultimate 3000 RSLC system using a BEH C18, 100 x 2.1 mm, 1.7 µm d_p column (Waters, Germany). Separation of 1 µL sample was achieved by a linear gradient from (A) H2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 600 µL/min and 45 °C. The gradient was 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 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 maXis 4G hr-ToF mass spectrometer (Bruker Daltonics, Germany) using the Apollo II ESI source. Mass spectra were acquired in centroid mode ranging from 150 – 2500 m/z at 2 Hz scan rate.

3.3.6 16S rRNA Gene and Phylogenetic Analysis

Extraction of the 16S rRNA gene was performed in representative microsclerodermin producing strains of *Sorangium, Jahnella* and *Chondromyces*. Sequences of other myxobacterial strains used in the analysis were obtained from GenBank. The 16S rRNA gene was amplified using a set of universal primers, and phylogenetic analysis was performed as described in a previous study, but using the MUSCLE alignment algorithm and Neighbor-Joining tree method (JC69) as implemented in the Geneious Pro program version 5.6.5.[35]

3.3.7 Genome Data

The *msc* gene cluster sequence was deposited in the GenBank with the accession no KF657738 for *S. cellulosum* So ce38 and accession no KF657739 for *Jahnella* sp. MSr9139.

3.4 Results and Discussion

3.4.1 Production of Microsclerodermins by Terrestrial Myxobacteria

In the course of our screening for bioactive natural products from myxobacteria, we observed antifungal activity in extracts from strain MSr9139, a newly isolated *Jahnella* species. Subsequent HPLC-based purification led to several fractions showing antifungal activity which contained compounds featuring an isotopic pattern typical for chlorination in MS analysis. Two compounds from these fractions could be assigned by their exact mass, fragmentation pattern, and retention time as pedein A (925.3493 *m/z*, $[M+H]^+$) and pedein B (891.3883 *m/z*, $[M+H]^+$), antifungal metabolites known from the myxobacterium *Chondromyces pediculatus* Cm p3.[25] Full structure elucidation was carried out for a compound with 869.3231 *m/z* obtained from another bioactive fraction, and the data unambiguously revealed this candidate as the known marine natural product microsclerodermin D (Table 1 and Figure S1). In addition to this, analysis of the MSr9139 extract led to the isolation and structure elucidation of the new derivative microsclerodermin L, differing from microsclerodermin D by an additional methoxy group, which is also reported for the pedein structure (Table 1 and Figure S3). Notably, the microsclerodermins are the first family of compounds found in the unexplored genus *Jahnella*, a member of the notable secondary metabolite producer myxobacterial family *Polyangiaceae*.[36]

Almost simultaneously, extracts from the myxobacterial strain *Sorangium cellulosum* So ce38 underwent biological profiling and HPLC fractionation, highlighting antifungal activity in the same chromatographic region as previously found from the MSr9139 extract. HPLC purification could narrow down the putatively active compounds to a candidate with 887.3934 *m/z*, and subsequent NMR analysis identified a peptide featuring the pyrrolidone moiety also known from microsclerodermins. NMR data revealed the presence of a new non-chlorinated derivative, microsclerodermin M (Table 1 and Figures 2, S4). It

shares the typical cyclic core structure with other microsclerodermins but features an unbranched side chain with three double bonds in conjugation to a phenyl moiety. Like the known microsclerodermins, the newly identified derivatives show potent activity against *Candida albicans* (microsclerodermin M, MIC 0.16 µg/mL; microsclerodermin L, MIC 18 µg/mL, microsclerodermin D, MIC 6.8 µg/mL). The stereochemistry of the isolated microsclerodermins was identified by acetonide formation and chemical degradation experiments followed by advanced Marfey analysis (see supporting information). It is identical to that reported for the microsclerodermins C – I and pedeins.[25,27,28]

Having discovered that myxobacteria from three different genera *Jahnella*, *Sorangium,* and *Chondromyces* are able to produce microsclerodermin congeners including even the exact same structure as previously described from two species of lithistid sponges (microsclerodermin D) was surprising for two reasons: examples of myxobacteria producing an identical scaffold also known from a phylogenetically distant organism are to date exceedingly rare (even when counting among the bacterial kingdom), and according to previous studies the secondary metabolite profiles from strains belonging to different myxobacterial genera usually exhibit little overlap.[6] In order to shed light on the occurrence of microsclerodermins within the myxobacteria, we conducted a search across high-resolution LC-MS data sets measured from almost 800 extracts, thus covering a sufficiently representative sample including most known myxobacterial taxa (Figure S20). On the basis of the evaluation of exact masses, isotope patterns and retention times we could identify a panel of 15 strains from the suborder *Sorangineae* (no single producer was found within the *Cystobacterineae*) as producers of microsclerodermins. Interestingly, our comprehensive LC-MS survey of myxobacterial secondary metabolomes revealed that producers of microsclerodermins form two mutually exclusive groups: one group comprises 11 *Sorangium* species producing solely the new microsclerodermin M, while the second group includes two strains of *Jahnella* sp. and two *Chondromyces* sp. that produce a variety of different derivatives: i.e. the "marine" microsclerodermin D in addition to the new microsclerodermin L and pedeins A/B (Figures S21 and S22). The various microsclerodermins differ in side chain, tryptophan modification and oxidation state at the pyrrolidone ring, whereas the peptidic core structure is always identical (Figure 1).

The fact that all microsclerodermins, irrespective of their origin, exhibit an identical macrocycle and in most cases even the same stereochemistry supports the idea of a shared biosynthetic origin or even a shared evolutionary ancestor. Moreover, the finding of a group of terrestrial myxobacteria producing exactly the same compound as found in lithistid sponges[27] (microsclerodermin D) fuels speculation about the actual biosynthetic origin of marine microsclerodermins. We herein propose that the marine microsclerodermins actually originate from a myxobacterium phylogenetically related to the *Sorangineae* suborder – possibly a yet uncultured species of the *Chondromyces*, *Jahnella,* or *Sorangium* taxa living in symbiosis with the sponges from which microsclerodermins were previously isolated.

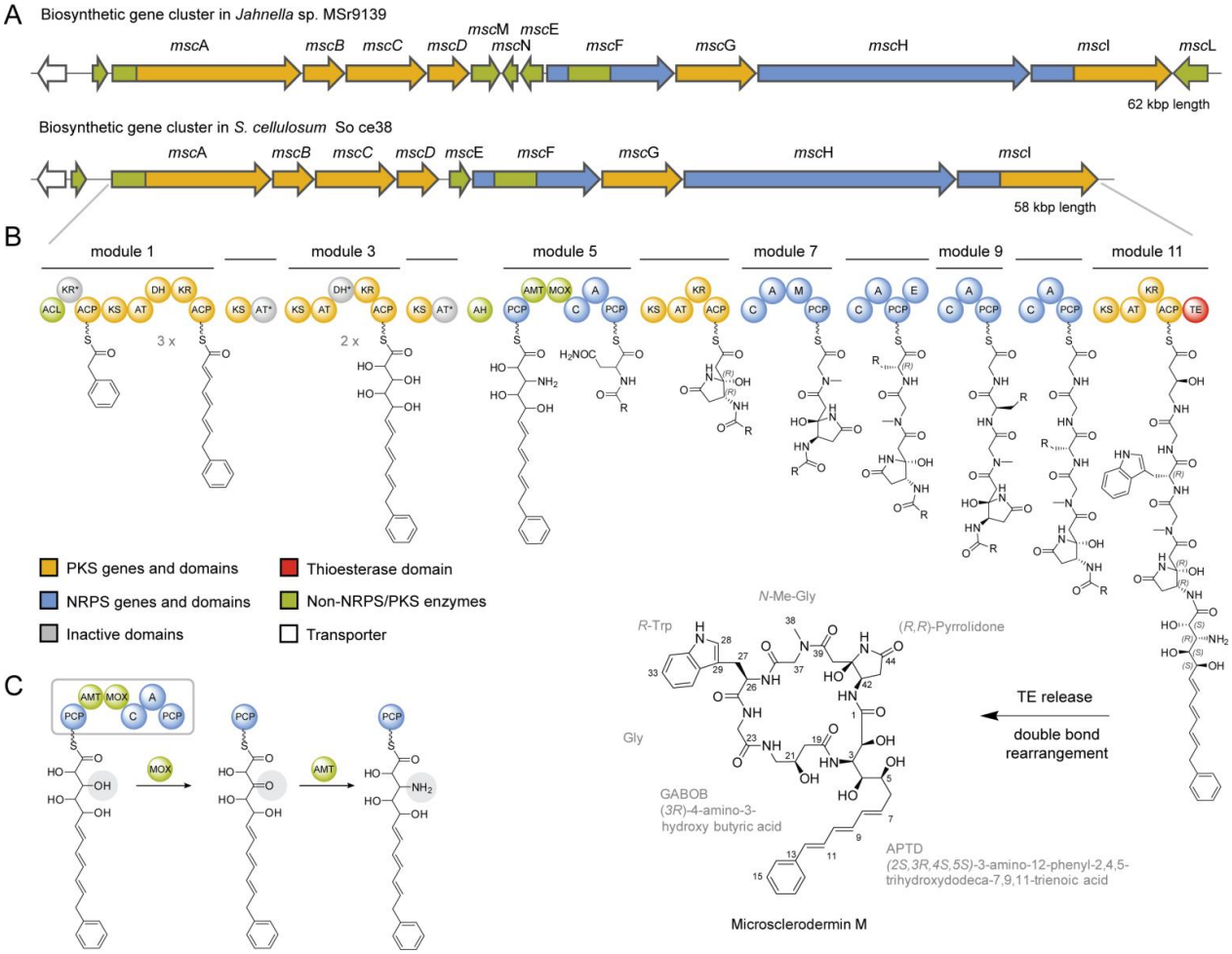


Figure 2: (A) Organization of the msc biosynthetic gene cluster in Jahnella sp. MSr9139 compared to Sorangium cellulosum So ce38. (B) Proposed biosynthetic route to microsclerodermin formation in So ce38. (C) Postulated biosynthetic steps leading to the amino group that is involved in macrolactam formation. A, adenylation domain; AMT, aminotransferase; ACP, acyl-carrier-protein domain; AT, acyltransferase; C, condensation domain; CoA-Lig, coenzyme A Ligase; DH, dehydratase; E, epimerase; KR, ketoreductase; KS, ketosynthase; MT, methyltransferase; MOX, monooxygenase; PCP, peptidyl-carrier-protein domain.

In coincidence with this hypothesis, phylogenetic studies of sponge metagenomes recently identified δ-proteobacteria in the sponge holobiont.[31] Indeed, the phylogenetic tree presented in the work of Simister *et al*. lists a clade containing nine myxobacterial species of terrestrial origin, including *Sorangium cellulosum* and *Chondromyces pediculatus*.[31] These data underpin our assumption that an evolutionary link exists between microsclerodermin biosynthesis in terrestrial and marine producers. Notably, 13 out of 15 producers identified by our LC-MS metabolome survey belong to the genera *Sorangium* or *Chondromyces*. In addition, the new *Jahnella* sp. MSr9139 was isolated from a soil sample collected from the same Philippine island where the sponge *Microscleroderma* was initially found.[27] This intriguing finding suggests that myxobacteria are possibly flushed to the ocean and adapted to an association with a sponge. The diversity and density of microbial flora present in sponges appears to be a good niche for a predator and proteo-bacteriolytic myxobacterium;[7] thus it is expected that in the future myxobacteria will be isolated from this underexplored source.

Access to myxobacterial producers holds a remarkable benefit, as these bacteria may be cultivated in large-scale fermentations, thereby allowing efficient production of the compounds of interest. Microsclerodermin M is produced at 12 mg/L in *S. cellulosum* So ce38 without optimization of growth conditions or genetic modification of the strain. Moreover, it allows us to investigate their biosynthesis, which has not been elucidated from any marine source to date. Thus, we set out to mine genome sequences of the new terrestrial producers for the presence of putative microsclerodermin biosynthetic pathways, using a retrobiosynthetic analysis as the starting point. The genome sequence of the strain *Sorangium cellulosum* So ce38, producer of the new microsclerodermin M, was already available from a previous study.[37] The newly isolated *Jahnella* sp. MSr9139 was selected for additional genome sequencing, as it produces the new microsclerodermin L in addition to known pedeins A and B and the "marine" microsclerodermin D.

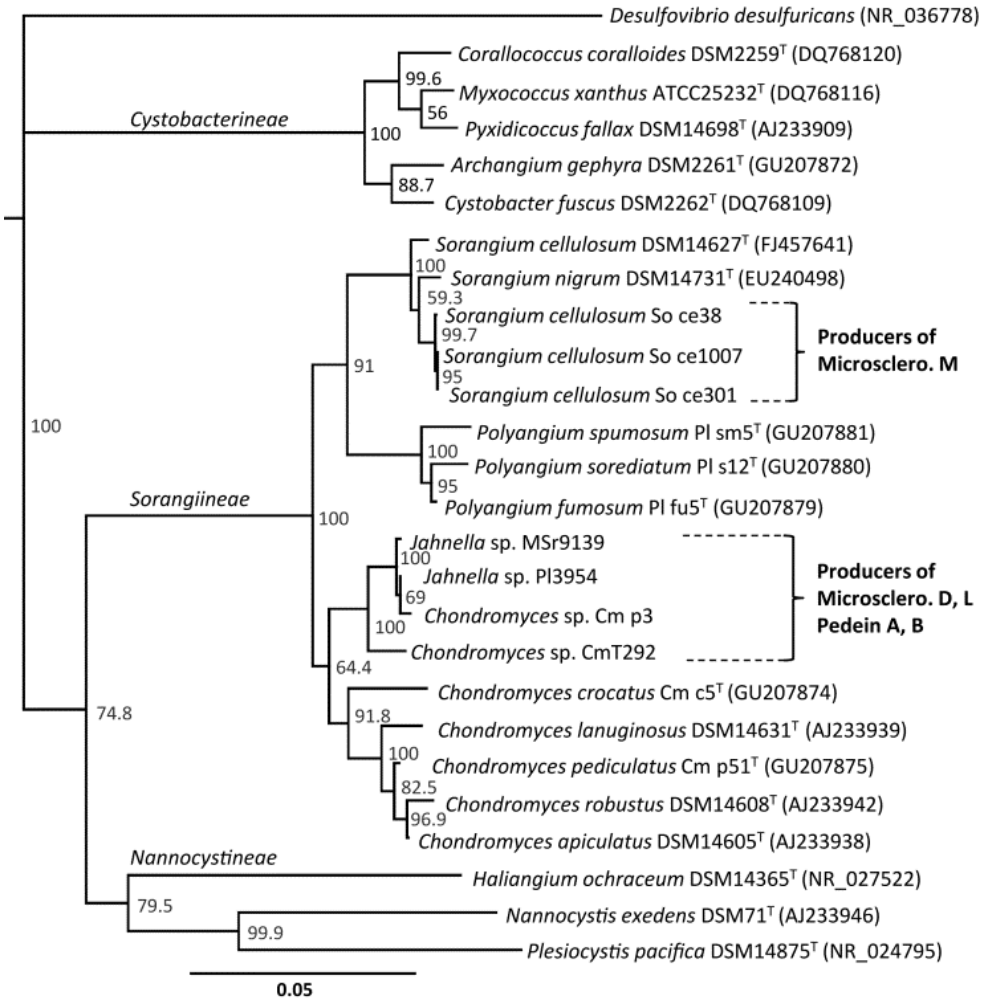


Figure 3: Neighbor-joining tree of myxobacteria inferred from 16S rRNA gene sequences showing the clades of microsclerodermin producing strains in suborder Sorangiineae. The numbers at branch points indicate percentage bootstrap support based on 1000 resamplings. GenBank accession numbers are indicated in parentheses. Bar = 0.05 substitutions per nucleotide position.

3.4.2 Microsclerodermin Biosynthetic Machinery

All microsclerodermins share the same cyclic peptide core but feature different lipophilic side chains and modifications of amino acid residues. On the basis of retrobiosynthetic considerations, the biosynthetic machinery for microsclerodermin formation was expected to consist of a multimodular PKS/NRPS system accompanied by a set of enzymes involved in side chain biosynthesis and post assembly line modification. The core PKS/NRPS modules should be conserved between producers, while enzymes involved in side chain biosynthesis and additional tailoring enzymes – responsible for modifications such as halogenation or oxidation of the pyrrolidone ring – should occur differentially between the two producer groups, as a consequence of evolutionary diversification of the microsclerodermin pathway.

Using *S. cellulosum* So ce38 and *Jahnella* sp. MSr9139 as representative strains from both microsclerodermin producer groups, we sought to identify microsclerodermin biosynthetic pathways in their genomes and subsequently elaborate on the molecular basis for the observed structural variations. The genome sequences of both strains were searched *in silico* for secondary metabolite gene clusters using the antiSMASH analysis pipeline.[38] The assignment of a matching candidate cluster to microsclerodermin biosynthesis was verified in So ce38 via targeted gene disruption by single crossover integration using biparental conjugation (Figure S24). Sequence comparison on the protein and nucleotide level revealed high similarity between gene clusters from both strains (Table 2), and comparison of operons permitted the tentative assignment of cluster boundaries. The microsclerodermin cluster spans a region of 58 kbp (74.7 % GC) in So ce38 and 62 kbp (72.4 % GC) in MSr9139, respectively. In both strains, genes encoding a major facilitator superfamily transporter (*mscK*) followed by a type II thioesterase (*mscJ*) are located upstream to *msc*A. The core biosynthetic assembly line covers five NRPS modules and three PKS modules encoded on genes *mscA* to *mscI*. An additional halogenase is encoded by *mscL* near the downstream boundary of the cluster in MSr9139 (Figure 2 A).

Microsclerodermin biosynthesis is initiated at the side chain to build up a phenyl group in conjugation to a double bond. An activated starter unit such as benzoyl-CoA or *trans*-cinnamoyl-CoA is usually recruited by the enzyme in such a case. However, a retrobiosynthetic proposal tells us that the observed double bond order of the side chain is likely different during biosynthesis (Figure 2, B). The biosynthetic logic requests the incorporation of C2 units, which is only possible if the double bonds are rearranged (Figure S23). A rearrangement of double bonds has already been reported for other natural products like bacillaen, rhizoxin, corallopyronin and ansamitocin where isomerization is likely catalyzed by a dehydratase domain.[39–42] The reason for isomerization of the microsclerodermin side chain remains elusive; however, it is supported by an energetic benefit of a conjugated π-system. On the basis of this hypothesis, the only suitable starter unit is phenylacetyl-CoA, which has already been reported for other natural product biosynthesis.[43] The incorporation of a phenylacetate starter unit was indeed verified by

feeding experiments using isotope-labeled precursors. Feeding ring-labeled $^{13}C_6$-L-phenylalanine resulted in a mass increase of 6 Da, whereas the fully labeled $^{15}N,^{13}C_9$-L-phenylalanine led to a mass shift of 8 Da, indicating the incorporation of two side chain carbons (Figure S27). Feeding d5-benzoic acid or d7-*trans*-cinnamic acid resulted in no mass increase (Figure S25). We can conclude that the α and β carbon atoms of phenylalanine - but not the carboxyl carbon - is incorporated into microsclerodermin. Elongation of the phenylacetate unit is catalyzed by modules MscA and MscC using three times malonate and two times 3-hydroxymalonate as extender units in an iterative manner. Modules 2 (MscB) and 4 (MscD) do only exhibit a combination of a functional KS domain attached to an inactive AT domain as identified by consensus sequence analysis, likely a relic of a former PKS complex.

Table 1: Proteins involved in microsclerodermin biosynthesis as identified in two myxobacterial strains.

	Sorangium cellulosum So ce38		*Jahnella* sp. MSr9139		
protein	length [aa]	domains and position in sequence	length [aa]	domains and position in sequence	identity [%]
MscA	3275	CoA-Lig (264-701), KR*(1034-1197), ACP (1348-1411), KS (1441-1828), AT (1976-2286), DH (2342-2505), KR' (2870-3047), ACP' (3149-3214)	3535	CoA-Lig (215-649), KR*(1032-1126), MT (1229-1509), ACP (1613-1676), KS (1712-2147), AT (2244-2555), DH (2610-2774), KR*' (3132-3309), ACP' (3411-3476)	65.4
MscB	870	KS (27-451), AT* (548-772)	878	KS (39-464), AT* (561-792)	73.4
MscC	1551	KS (36-460), AT (557-859), DH* (956-1076), KR (1158-1336), ACP (1436-1505)	1549	KS (39-464), AT (561-865), DH* (956-1076), KR (1157-1335), ACP (1436-1499)	75.0
MscD	900	KS (36-461), AT* (565-678)	848	KS (36-461), AT* (564-675)	72.0
MscE	446	Putative amidohydrolase	386	Putative amidohydrolase	82.9
MscF	2273	PCP (27-98), AMT (329-660), MOX (828-1127), C (1185-1529), A (1673-2082), PCP' (2169-2237)	2189	PCP (4-75), AMT (280-614), MOX (758-1057), C (1105-1397), A (1594-2000), PCP' (2088-2149)	76.5
MscG	1548	KS (14-439), AT (534-828), KR (1156-1330), ACP (1441-1509)	1511	KS (14-439), AT (531-850), KR (1136-1312), ACP (1404-1472)	72.1
MscH	4141	C (76-377), A (564-966), MT (1037-1256), PCP (1469-1531), C' (1554-1850), A' (2037-2426), PCP' (2515-2574), E (2591-2905), C'' (3074-3374), A'' (3558-3967), PCP'' (4054-4121)	4106	C (48-346), A (534-936), MT (1007-1225), PCP (1442-1505), C' (1526-1827), A' (2013-2405), PCP' (2492-2551), E (2568-2872), C'' (3043-3343), A'' (3527-3932), PCP'' (4019-4083)	78.8
MscI	2904	C (48-346), A (533-936), PCP (1023-1087), KS (1111-1535), AT (1638-1936), KR (2266-2465), ACP (2549-2612), TE (2634-2888)	2945	C (77-375), A (563-965), PCP (1053-1116), KS (1150-1573), AT (1676-1970), KR (2309-2509), ACP (2597-2660), TE (2683-2945)	77.8
MscJ	257	thioesterase type II	263	thioesterase type II	43.5
MscK	415	major facilitator superfamily (MFS) transporter	450	major facilitator superfamily (MFS) transporter	24.5
MscL	-	-	535	Tryptophan halogenase	-
MscM	-	-	438	Fe(II)/α-ketoglutarate dependent oxygenase	-
MscN	-	-	277	methyltransferase	-

A, adenylation domain; AMT, aminotransferase; ACP, acyl-carrier-protein domain; AT, acyltransferase; C, condensation domain; CoA-Lig, coenzyme A Ligase; DH, dehydratase; E, epimerase; KR, ketoreductase; KS, ketosynthase; MT, methyltransferase; MOX, monooxygenase; PCP, peptidyl-carrier-protein domain. * inactive domain

The PKS-derived unit is forwarded to the first PCP domain of module MscF. This module harbors two additional domains of rather uncommon type showing high homology to the amino transferase family (AMT) and to the monooxygenase family, both located downstream to the PCP domain. A biosynthetic proposal to account for this domain order is based on oxidation of the β-hydroxyl group of the bound intermediate to the respective β-keto functionality followed by conversion to a β-amino moiety that undergoes macrocyclization (Figure 2, C). The use of an aminotransferase is known from other natural product biosynthetic pathways, however, not in combination with the initial oxidation step.[44]

Thereafter, biosynthesis continues with a set of NRPS- and PKS-based reaction cycles. Analysis of the A domain specificities *in silico* is consistent with the amino acids incorporated.[45] We propose the uncommon pyrrolidone moiety is built up by asparagine cyclization. To the best of our knowledge, such an asparagine-derived pyrrolidone system is only found in microsclerodermins and koshikamides, a natural product that was isolated from a *Theonella* species.[46] Indeed, the A domain of MscF is specific for asparagine activation and we did a feeding experiment with fully labeled $^{15}N_2$-$^{13}C_4$-L-asparagine to prove this biosynthetic step. We observed a mass increase of 6 Da according to the incorporation of all carbon and nitrogen atoms of asparagine into the compound (Figures S26, S28). On the basis of this result, a plausible biosynthetic hypothesis requires the nucleophilic attack of the asparagine side chain to the backbone carbonyl atom. A suitable mechanism is known from the intein-mediated peptide cleavage, where intein initiates an intra-molecular asparagine cyclization, notwithstanding the poor reactivity of the side chain's amide.[47] In microsclerodermin biosynthesis, this reaction likely is accompanied by an inversion of the stereochemistry at the α-carbon of the former (*S)*-asparagine. The relative configuration of the pyrrolidone ring was identified by NOE correlations, whereas the absolute (*R,R*)-configuration is derived from degradation experiments (see supporting information). The stereochemistry at this position is thereby identical with that of microsclerodermins C – I. The protein MscE is most likely responsible for the cyclization step, as it is found in both microsclerodermin clusters and shows similarity to the amidohydrolase class, a fairly promiscuous enzyme family able to act on a variety of substrates. However, the exact mechanism involved in this biosynthetic step remains elusive at present.

The forthcoming NRPS modules correspond to the observed structure of microsclerodermin in terms of domain order and predicted substrate specificity (Figure 2, B). For both new microsclerodermins an *R*-configured tryptophan was identified by means of the advanced Marfey method, which is in agreement with the epimerization domain found in module 8. The (3*R*)-configuration of the γ-amino butyric acid (GABA) subunit was identified by the same technique (see

Figures S13, S14). Both stereogenic centers have the same configuration as identified in all microsclerodermins so far.

3.4.3 Genetic Basis for the Structural Diversity of Microsclerodermins

The derivatives found in *Jahnella* sp. MSr9139 feature side chains with either one or two double bonds while the side chain in So ce38 comprises strictly three double bonds**.** As the number of PKS modules encoded in the gene cluster does not match the number of required elongation cycles, an iterative function of the type I PKS subunits MscA and MscC as described for the stigmatellin megasynthase may explain this finding.[48] The KS domains of each module are highly identical for both strains and do not comprise any of the postulated sequence-based identifiers of iterative KS domains.[49] Nevertheless, MscB is grouping with iterative KS domains in a phylogenetic analysis (Figure S11). Comparing the entire module MscA of both clusters revealed the insertion of an additional methyl transferase-like domain into the first part of the protein in MSr9139. This domain is likely inactive on the basis of *in silico* analysis as judged by the presence of a corrupted SAM-binding motif (Table S13).[50] Currently, we cannot rule out the possibility that the presence of this additional methyl transferase may influence the iteration process within MscA. In addition to this difference, there is no obvious reason why biosynthesis in So ce38 results in a triene whereas MSr9139 is less strict in iteration. As another hint for a shared origin of the biosynthetic cluster, some of the sponge-derived derivatives exhibit a methyl-branched side chain which could be attributed to this methyl transferase being active in some of the marine producers. Eventually, the variable side chains of the microsclerodermin family are in agreement with the alternating PKS functionality. Halogenation of tryptophan as well as oxidation of the pyrrolidone ring is catalyzed by tailoring enzymes. The halogenase MscL is located downstream to the cluster in MSr9139 and is responsible for chlorination of the tryptophan. It shows 32 % identity on a protein level to a tryptophan halogenase from a Streptomyces species (PDB entry 2WET_A). There is no analogue of MscL found in So ce38, which is in agreement with the absence of chlorinated products in this strain. Supplementing KBr or NaBr to the MSr9139 cultivation broth led to the production of brominated microsclerodermins on the basis of LC-MS analysis. Another difference is the inter-region between the main PKS and NRPS parts. In MSr9139 two additional proteins are found in this region. MscN is a member of the SAM-dependent methyl transferase family, and MscM shows homology to Fe(II)/α-ketoglutarate-dependent dioxygenases. On the basis of the structures produced by MSr9139, we conclude that MscM and MscN are responsible for oxidation and methylation of the pyrrolidone ring, respectively. Modifications at the tryptophan as known from some marine-derived microsclerodermins were not observed in this study. Such modifications are attributed to promiscuous acting enzymes that could be related to the producer strain or even to enzymes related to some sponge symbiont.

3.5 Conclusion

The discovery of microsclerodermins/pedeins from several myxobacteria represents one of the few findings of identical compounds from marine organisms and terrestrial bacteria reported to date. This study thus strengthens the notion that certain natural products, which have been isolated from marine sources such as sponges or other invertebrates, actually originate from associated microbes. Notably, the identification of microsclerodermin-producing myxobacteria provides meaningful hints for future attempts to isolate the symbiotic microbe. This knowledge is considered particularly helpful because isolation success in many cases critically depends on methods well adapted to the requirements of the genus targeted for isolation, especially when aimed at the rather challenging isolation of slow-growing myxobacteria. Availability of an alternative microbial producer as a sustainable source is an advantage for realizing the potential of a marine natural product for therapeutic applications. Moreover, the myxobacterial producers come along with additional chemical diversity and are amenable to genetic manipulation, as demonstrated in this study. Finally, the identification of two slightly different microsclerodermin biosynthetic gene clusters from two myxobacteria allowed us to establish a conclusive model for microsclerodermin biosynthesis and provided insights into the molecular basis for structural diversity within this compound family. A detailed understanding of microsclerodermin biosynthesis is an important prerequisite for any future efforts toward engineering the pathway for yield improvement or for the production of new derivatives: whether in the native producer, by heterologous expression, or by using synthetic biology approaches.

3.6 Supporting Information

3.6.1 Structure elucidation

3.6.1.1 NMR data of Microsclerodermin D

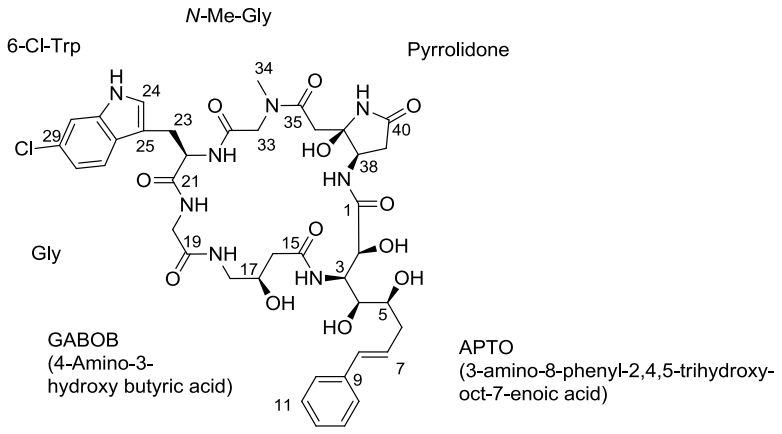


Figure S1: Structure and atom numbering of microsclerodermin D. Compound isolated from marine sponges of the genera Microscleroderma sp. and Theonella sp. and the myxobacterium Jahnella sp. MSr9139.

Table S1: NMR chemical shifts of microsclerodermin D.

Amino acid	Assignment	δ_C [a]	δ_H (mult., *J* in Hz) [b]	HMBC	ROESY
APTO	1	172.3			
	2	69.2	4.39 (d, 3.2)		
	3	53	4.13 (d, 11.3)		
	4	69.5	3.32 (m)		
	5	68.8	3.57 (m)		
	6	36.3	2.35 (m)		
	7	128.0	6.24 (m)		
	8	130.5	6.40 (d, 15.8)		
	9	137.2			
	10, 14	125.4	7.35 (d, 7.3)		
	11, 13	128.2	7.29 (t, 7.6)		
	12	126.5	7.19 (t, 7.3)		
	OH-2		6.58 (brs)		
	OH-4		4.17 (m)		
	OH-5		4.36 (d, 4.5)		
	NH-3		7.47 (d, 6.0)		
GABOB	15	172.3			
	16	40.7	2.43 (d, 14.0)		
			2.15 (d, 14.0)		
	17	66.8	3.72 (m)		
	18	44.7	3.39 (m)		
			2.61 (m)		
	OH-17		4.88 (d, 4.8)		
	NH-18		7.50 (m)		
Gly	19	168.5			
	20	42.3	3.75 (d, 7.0)		
			3.34 (d, 7.0)		
	NH-20		8.54 (t, 6.0)		

6-Cl-Trp	21	171.5	
	22	55.1	4.17 (m)
	23	25.8	3.10 (dd, 5.7, 14.7)
			2.98 (dd, 5.7, 14.7)
	24	124.6	7.26 (d, 1.8)
	25	109.5	
	26	125.6	
	27	119.4	7.52 (d, 8.3)
	28	118.4	7.00 (dd, 1.5, 8.3)
	29	c	
	30	110.7	7.37 (s)
	31	136.1	
	NH-22		8.64 (d, 4.3)
	NH-24		11.04 (brs)
N-Me-Gly	32	169.9	
	33	49.5	4.08 (d, 16.3)
			3.84 (d, 16.3)
	34	36.2	2.93 (s)
Pyrrolidone	35	170.2	
	36	38.6	2.84 (d, 17.1)
			2.69 (d, 17.1)
	37	85.2	
	38	50.3	4.47 (m)
	39	35.0	2.27 (m)
	40	172.3	
	NH-37		7.96 (brs)
	NH-38		7.56 (m)
	OH-37		c

[a] Recorded at 175 MHz, referenced to residual solvent DMSO-d_6 at 39.51 ppm.

[b] Recorded at 700 MHz, referenced to residual solvent DMSO-d_6 at 2.50 ppm.

[c] Not observed. [ov] overlapping signals.

3.6.1.2 NMR data of Dehydromicrosclerodermin D - acetonide

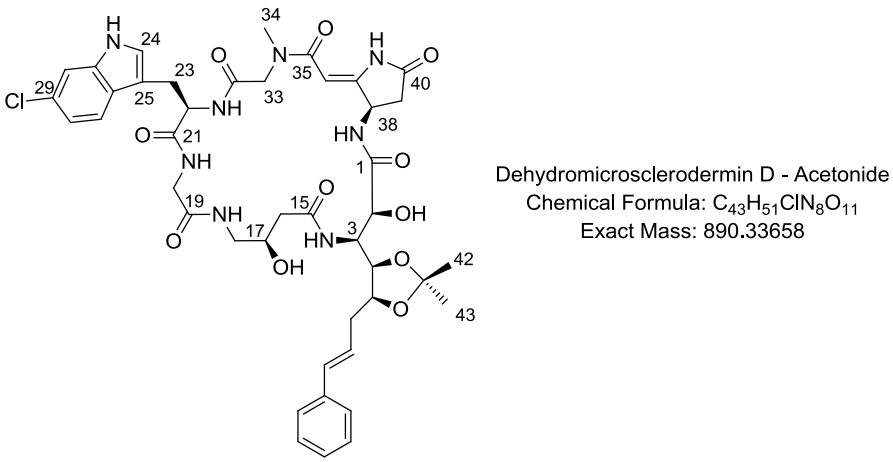


Figure S2: Structure and atom numbering of the acetonide of dehydromicrosclerodermin D. Derivative is based on microsclerodermin D as isolated from the myxobacterium Jahnella sp. MSr9139.

Table S2: NMR chemical shifts of dehydromicrosclerodermin D - acetonide.

Amino acid	Assignment	δ_C [a]	δ_H (mult., *J* in Hz) [b]
APTO	1	172.3	
	2	70.1	4.19 (d, 4.2)
	3	55.4	4.11 (d, 10.0)
	4	77.3	3.74 (ov)
	5	78.7	4.02 (m)
	6	37.1	2.32 (m) 2.55 (m)
	7	126.5	6.24 (m)
	8	131.3	6.43 (d, 16.0)
	9	c	
	10, 14	125.4	7.38 (m)
	11, 13	128.2	7.30 (t, 7.6)
	12	126.5	7.20 (t, 7.2)
	OH-2		6.23 (ov)
	NH-3		7.09 (ov)
GABOB	15	c	
	16	41.1	2.22 (d, 14.8)
			2.01 (m)
	17	65.6	3.89 (m)
	18	44.7	3.23 (m)
			2.82 (m)
	OH-17		5.10 (d, 4.5)
	NH-18		7.35 (t, 6.5)
Gly	19	c	
	20	42.6	3.58 (m))
	NH-20		8.38 (t, 6.5)
6-Cl-Trp	21	c	
	22	55.2	4.11 (m)
	23	25.4	3.19 (ov)
			3.01 (dd, 5.1, 10.0)
	24	124.6	7.26 (d, 1.8)
	25	109.5	
	26	125.6	
	27	119.4	7.52 (d, 8.3)
	28	118.4	7.00 (dd, 1.5, 8.3)
	29	c	
	30	110.7	7.37 (s)
	31	136.1	
	NH-22		8.77 (d, 5.0)
	NH-24		11.04 (brs)
N-Me-Gly	32	c	
	33	50.5	3.44 (d, 15.7)
			4.53 (d, 15.7)
	34	36.8	3.06 (s)
Pyrrolidone	35	c	
	36	88.0	5.25 (s)
	37	c	
	38	45.4	5.21 (m)
	39	33.7	2.44 (m) 2.74 (m)
	40	c	
	NH-37		c
	NH-38		8.36 (d, 8.8)

[a] Recorded at 175 MHz, referenced to residual solvent DMSO-d_6 at 39.51 ppm.

Recorded at 700 MHz, referenced to residual solvent DMSO-d_6 at 2.50 ppm.

[c] Not observed. [ov] overlapping signals.

3.6.1.3 NMR data of Microsclerodermin L

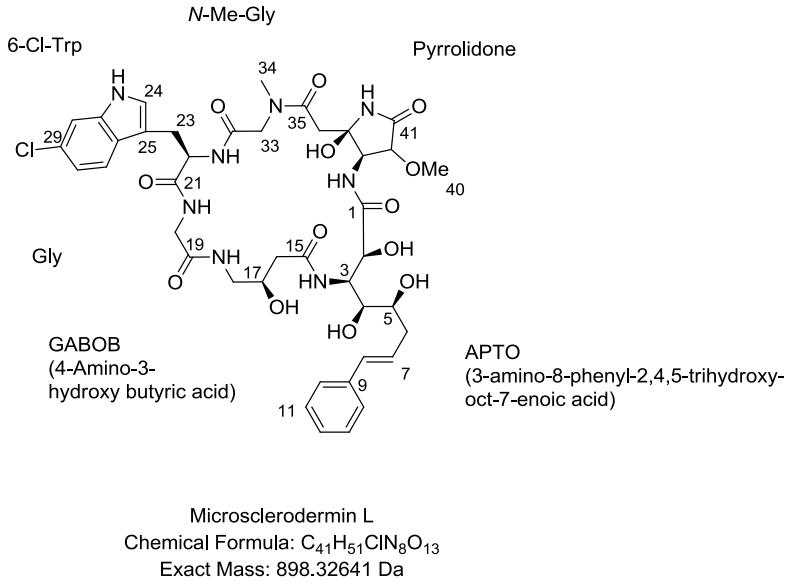


Figure S3: Structure and atom numbering of microsclerodermin L. Compound isolated from the myxobacterium Jahnella sp. MSr9139.

Table S3: NMR chemical shifts of microsclerodermin L.

Amino acid	Assignment	δ_C [a]	δ_H (mult., *J* in Hz) [b]	HMBC	ROESY
APTO	1	172.9			
	2	69.2	4.43 (brs)	1	
	3	53	4.19 (d, 11.3)	4	4
	4	69.7	3.32 (m)		
	5	68.7	3.59 (q, 6.5)		
	6	36.6	2.38 (m)	4, 5, 7, 8	
			2.33 (m)		
	7	128	6.26 (m)	6, 9	
	8	130.5	6.40 (d, 15.8)	6, 9, 10, 14	6
	9	137.2			
	10, 14	125.6	7.36 (d, 7.0)	8, 12	7, 8
	11, 13	128.3	7.29 (t, 7.6)	9, 10, 14	
	12	126	7.19 (t, 7.4)	10, 11, 13, 14	
	OH-2		6.57 (brs)		
	OH-4		4.85 (d, 4.9)		
	OH-5		4.30 (brs)		
	NH-3		7.42 (brs)		16
GABOB	15	172.1			
	16	40.6	2.43 (d, 14.0)	15	
			2.12 (d, 14.0)		
	17	66.6	3.77 (m)	16, 18	OH-17
	18	44.8	3.33 (m)	17	
			2.65 (m)		
	OH-17		4.85 (brs)		
	NH-18		7.43 (brs)		20
Gly	19	168.5			
	20	42.5	3.72 (d, 7.0)	19, 21	
			3.36 (d, 7.0)		
	NH-20		8.51 (t, 6.0)		22
6-Cl-Trp	21	171.6			
	22	55.2	4.17 (m)		NH-20
	23	25.8	3.10 (dd, 5.7, 15.0)	21, 22, 24, 25, 26	
			2.98 (dd, 5.7, 15.0)		

	24	124.7	7.26 (d, 2.0)	25, 31, 32	
	25	109.3			
	26	125.5			
	27	119.4	7.52 (d, 8.5)	31	
	28	118.5	7.00 (dd, 1.8, 8.5)	30, 26	
	29	c			
	30	110.7	7.37 (s)	27	
	31	136.3			
	NH-22		8.68 (d, 4.3)		22, 23, 24, 33
	NH-24		11.03 (d, 2.0)	24, 25, 31, 26	24, 30
N-Me-Gly	32	172.2			
	33	49.6	4.28 (d, 16.4)	32, 34	
			3.69 (d, 16.4)		
	34	35.9	2.92 (s)	33, 35	
Pyrrolidone	35	170.4			
	36	38.6	2.83 (d, 16.9)	35, 37	38
			2.61 (d, 16.9)		
	37	82.3			
	38	55.8	4.44 (d, 9.0)	39	NH-38
	39	79.0	4.04 (d, 8.8)	40, 41	NH-38
	40	56.6	3.39 (s)	39	
	41	171.0			
	NH-37		8.18 (brs)	37, 38, 39	
	NH-38		7.61 (d, 9.6)		2, 39
	OH-37		c		

[a] Recorded at 175 MHz, referenced to residual solvent DMSO-d_6 at 39.51 ppm.

Recorded at 700 MHz, referenced to residual solvent DMSO-d_6 at 2.50 ppm.

[c] Not observed. [ov] overlapping signals.

3.6.1.4 NMR data of Microsclerodermin M

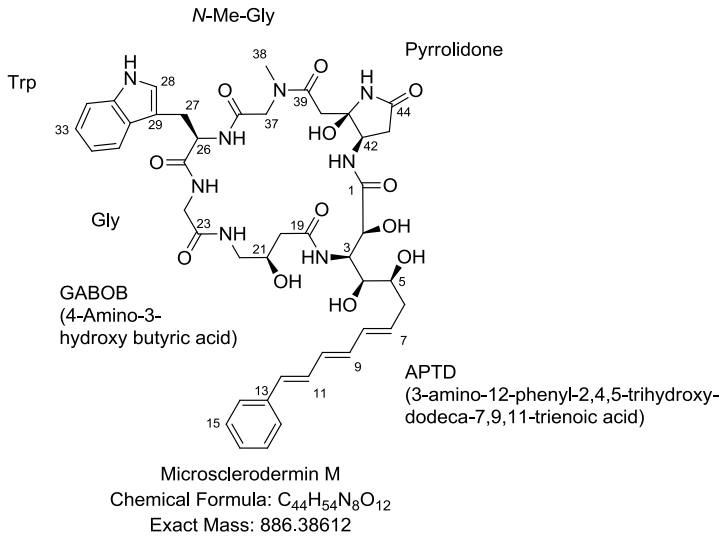


Figure S4: Structure and atom numbering of microsclerodermin M. Compound isolated from the myxobacterium Sorangium cellulosum So ce38.

Table S4: NMR chemical shifts of microsclerodermin M.

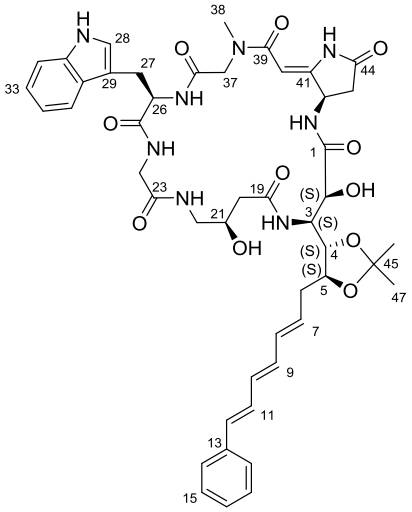
Amino acid	Assignment	δ_C [a]	δ_H (mult., *J* in Hz) [b]	HMBC	ROESY
APTD	1	172.5			
	2	69.3	4.38 (d, 5.4)	3	3, OH-2
	3	53.0	4.11 (ov)	4, 5, 19	OH-5, NH-3
	4	69.8	3.27 (ov)	2, 3	3, 7, OH-2, OH-3
	5	68.8	3.52 (ov)	4, 6	7, 8
	6	36.4	2.28 (ov)	4, 5, 7, 8	4, 5, 7, 8
	7	133.0	5.74 (dt, 15.0, 7.4)	6, 9	4, 5, 6, 9
	8	131.7	6.17 (dd, 10.0, 15.1)	9, 10	6, 11
	9	133.7	6.34 (ov)	7, 8, 10, 11	7, 11
	10	130.7	6.33 (ov)	8, 11, 12	12
	11	129.2	6.92 (dd, 9.6, 15.6)	9, 10, 13	9, 18
	12	131.3	6.56 (d, 15.6)	10, 14, 18, 13	10, 14
	13	137.0			
	14, 18	125.7	7.44 (d, 7.6)	12, 16	11, 12
	15, 17	128.4	7.31 (d, 8.3)	14, 18	14
	16	127.1	7.21 (d, 7.1)	14, 15, 17, 18	14, 18
	OH-2		6.25 (d, 5.6)	2, 3	2, 4, NH-42
	OH-4		4.34 (d, 4.7)	5, 6	3, 5
	OH-5		4.52 (d, 9.0)	3, 4, 5	3, 4
	NH-3		7.43 (d, 12.6)	2, 3	4, 5, 20, OH-21
GABOB	19	172.6			
	20	40.7	2.44 (d, 14.0)	19, 21, 22	
			2.15 (ov)		NH-3
	21	66.9	3.73 (ov)	21, 23	20, 22, OH-21, NH-22
	22	44.8	3.37 (ov)		21, OH-21
			2.64 (ov)		
	NH-22		7.49 (d, 6.7)		20, 21, 22
	OH-21		4.86 (d, 5.0)	20, 21, 22	20, 21, 22, NH-3
Gly	23	168.7			
	24	42.4	3.76 (d, 7.6)	23, 25	NH-24
			3.36 (d, 5.7)		
	NH-24		8.56 (t, 5.7)		24, 26, NH-22
Trp	25	171.7			
	26	55.2	4.19 (m)	25, 27, 29, 36	27, 30, NH-24
	27	26.2	3.11 (dd, 5.6, 14.6)	25, 26, 29	26, 28, 30, NH-26
			2.98 (dd, 5.6, 14.6)		
	28	123.5	7.21 (d, 6.6)	29, 34	26, 27, NH-26, NH-28
	29	109.6			
	30	126.5			
	31	118.0	7.52 (d, 7.7)	29, 32	26, 27, 32
	32	118.1	6.98 (t, 7.3)	33, 34	31, 34
	33	120.8	7.06 (d, 7.3)	30, 31, 35	31, 34
	34	111.1	7.32 (d, 7.9)	30, 32, 35	32
	35	135.9			
	NH-28		10.87 (brs)	28, 29, 35, 30	28, 34
	NH-26		8.65 (d, 4.2)	26, 27	26, 27, 28, 37
N-Me-Gly	36	169.9			
	37	49.6	4.10 (d, 15.4)	39, 36	38, NH-26
			3.82 (d, 15.4)		
	38	36.2	2.94 (s)	39	37
Pyrrolidone	39	170.4			
	40	38.7	2.85 (d, 17.0)	39, 41, 42	42, OH-41, NH-41
			2.70 (d, 17.0)		
	41	85.5			
	42	50.4	4.47 (ov)	40, 42, 43	40
	43	34.9	2.28 (ov)	41, 44	42, OH-41, NH-42
	44	172.9			
	OH-41		6.07 (s)	40, 42	40, 43, NH-41, NH-42
	NH-41		7.97 (s)	41, 42, 43	40, 42, OH-41
	NH-42		7.54 (ov)	42	42, 43, OH-41

[a] Recorded at 175 MHz, referenced to residual solvent DMSO-d_6 at 39.51 ppm.

Recorded at 700 MHz, referenced to residual solvent DMSO-d_6 at 2.50 ppm.

[c] Not observed. [ov] overlapping signals.

3.6.1.5 NMR data of Dehydromicrosclerodermin M - acetonide



Dehydromicrosclerodermin M
Acetonide
Chemical Formula: $C_{47}H_{56}N_8O_{11}$
Exact Mass: 908.40685

Figure S5: Structure and atom numbering of the acetonide of dehydromicrosclerodermin M. Derivative is based on microsclerodermin M as isolated from the myxobacterium Sorangium cellulosum So ce38.

Table S5: NMR chemical shifts of microsclerodermin M - acetonide.

Amino acid	Assignment	δ_C [a]	δ_H (mult., *J* in Hz) [b]
APTD	1	172.5	
	2	69.9	4.17 (ov)
	3	53.1	4.28 (m)
	4	77.0	3.69 (dd, 7.0, 9.9)
	5	78.6	3.96 (m)
	6	36.7	2.47, 2.24 (ov)
	7	133.0	5.75 (m)
	8	131.7	6.17 (dd, 9.7, 15.0)
	9	133.7	6.34 (ov)
	10	130.7	6.33 (ov)
	11	129.2	6.92 (dd, 9.6, 15.6)
	12	131.3	6.56 (d, 15.6)
	13	137.0	
	14, 18	125.7	7.44 (d, 7.6)
	15, 17	128.4	7.31 (ov)
	16	127.1	7.21 (t, 7.3)
	45	*c*	
	46	26.6	1.35 (s)
	47	26.6	1.35 (s)
	OH-2		6.25
	NH-3		7.43
GABOB	19	172.6	
	20	40.3	2.22 (ov)
			2.00 (dd, 7.7, 10.8)
	21	65.2	3.87 (m)
	22	44.8	3.23 (ov)
			2.82 (m)
	NH-22		7.37

	OH-21		4.86
Gly	23	168.7	
	24	41.9	3.59 (m)
	NH-24		8.40
Trp	25	171.7	
	26	54.8	4.14 (m)
	27	25.1	3.21 (dd, 3.3, 10.3)
			3.00 (dd, 3.3, 10.3)
	28	123.5	7.21 (d, 5.3)
	29	109.6	
	30	126.5	
	31	118.0	7.52 (d, 7.7)
	32	118.1	6.98 (t, 7.3)
	33	120.8	7.06 (t, 7.3)
	34	111.1	7.32 (d, 7.9)
	35	135.9	
	NH-28		10.87 (brs)
	NH-26		8.65
N-Me-Gly	36	169.9	
	37	49.7	4.54 (d, 11.5)
			3.43 (ov)
	38	36.4	3.05 (s)
Pyrrolidone	39	170.4	
	40	87.3	5.23 (ov)
	41	*c*	
	42	50.4	3.69 (m)
	43	33.1	2.45, 2.75 (ov)
	44	172.9	
	NH-41		8.35
	NH-42		*c*

[a] Recorded at 175 MHz, referenced to residual solvent DMSO-d_6 at 39.51 ppm. Recorded at 700 MHz, referenced to residual solvent DMSO-d_6 at 2.50 ppm. [c] Not observed. [ov] overlapping signals.

3.6.2 Analytical data for microsclerodermins and pedeins

3.6.2.1 Microsclerodermin D

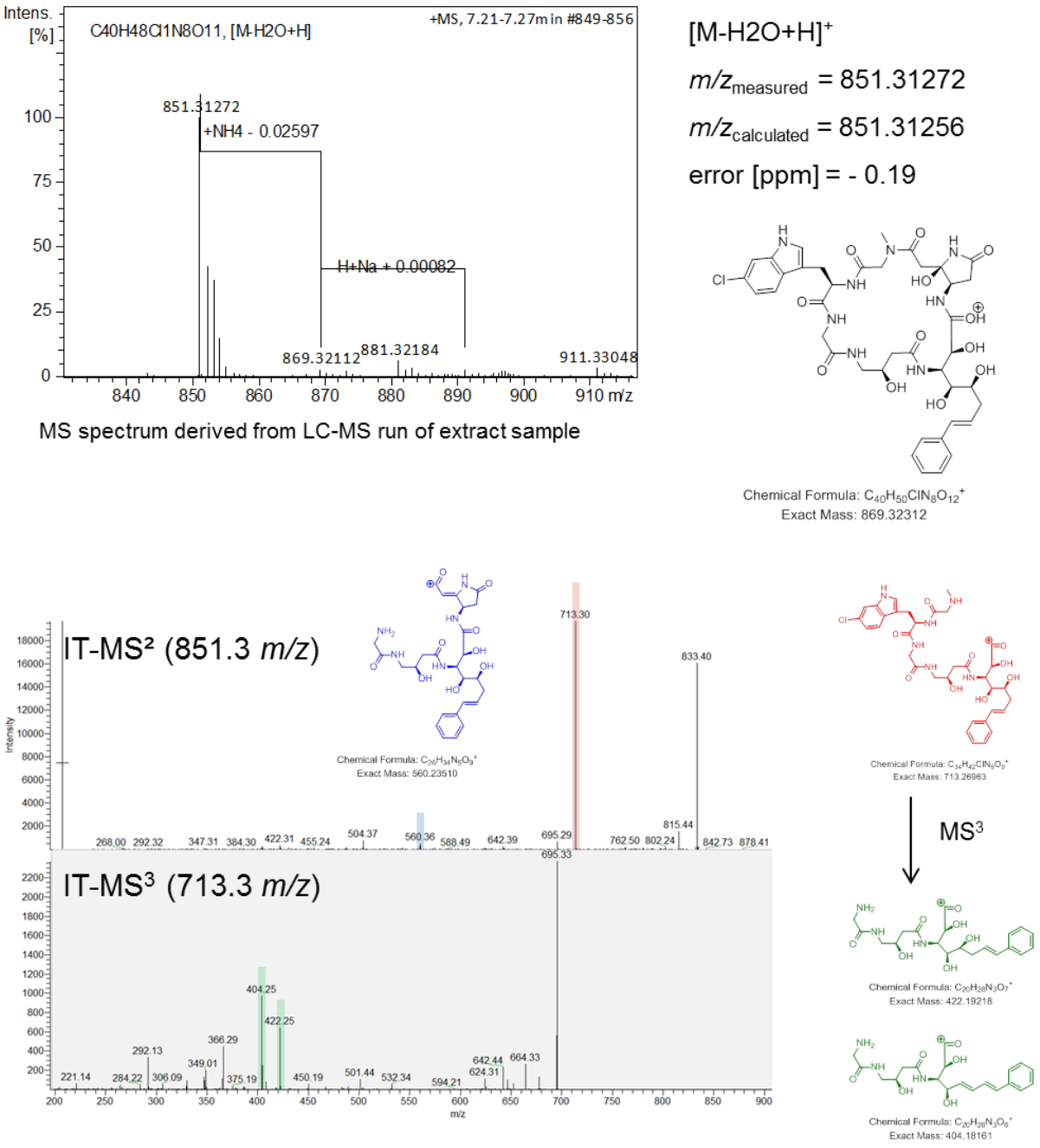


Figure S6: Accurate m/z measurement and CID fragmentation pattern of microsclerodermin D.

3.6.2.2 Microsclerodermin L

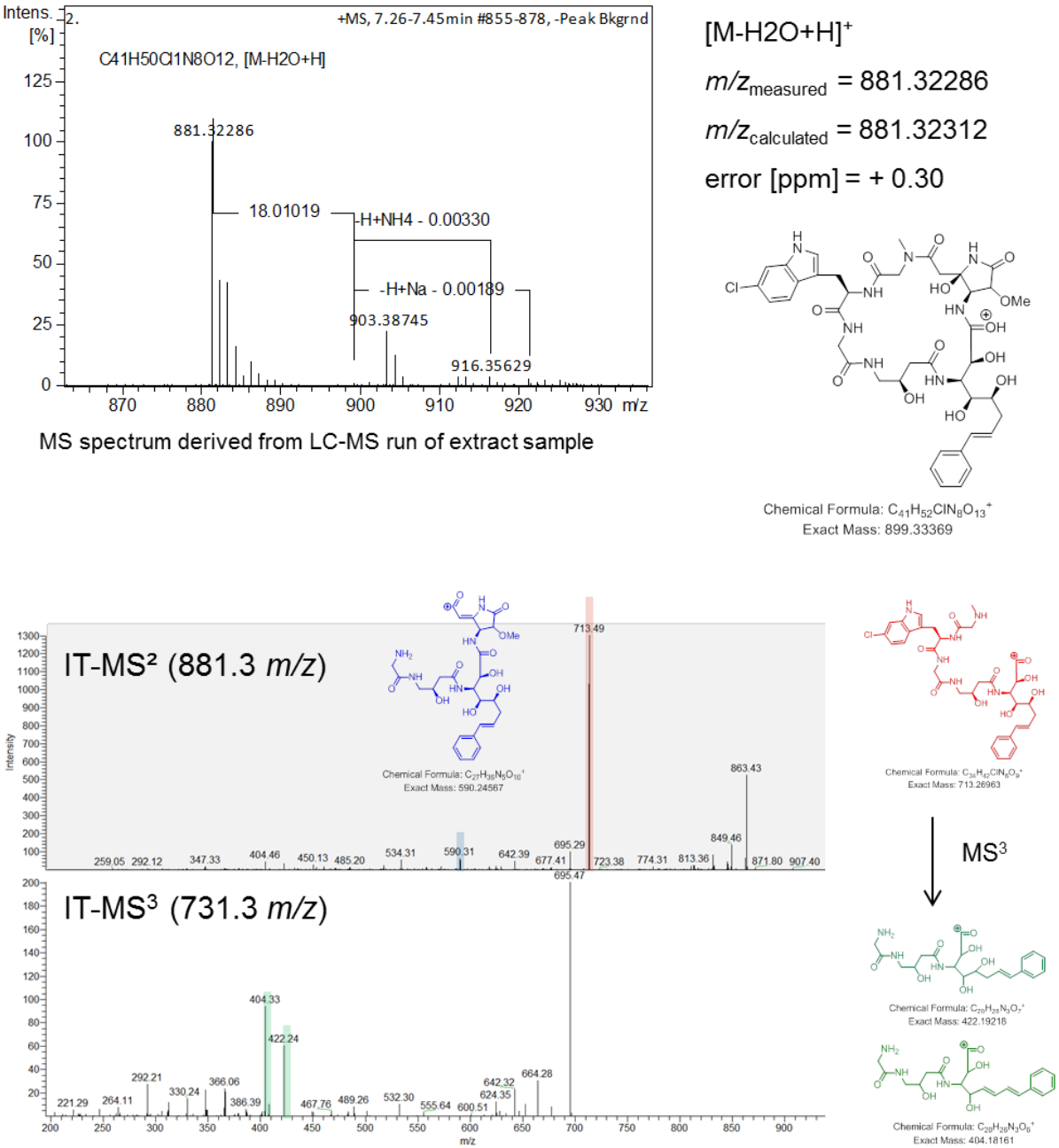
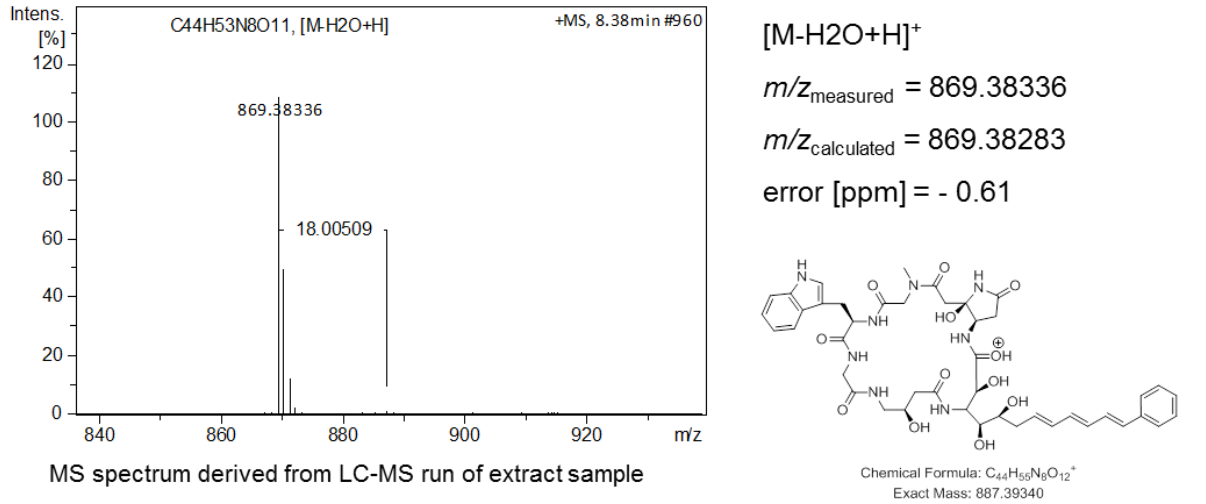


Figure S7: Accurate m/z measurement and CID fragmentation pattern of microsclerodermin L.

3.6.2.3 Microsclerodermin M



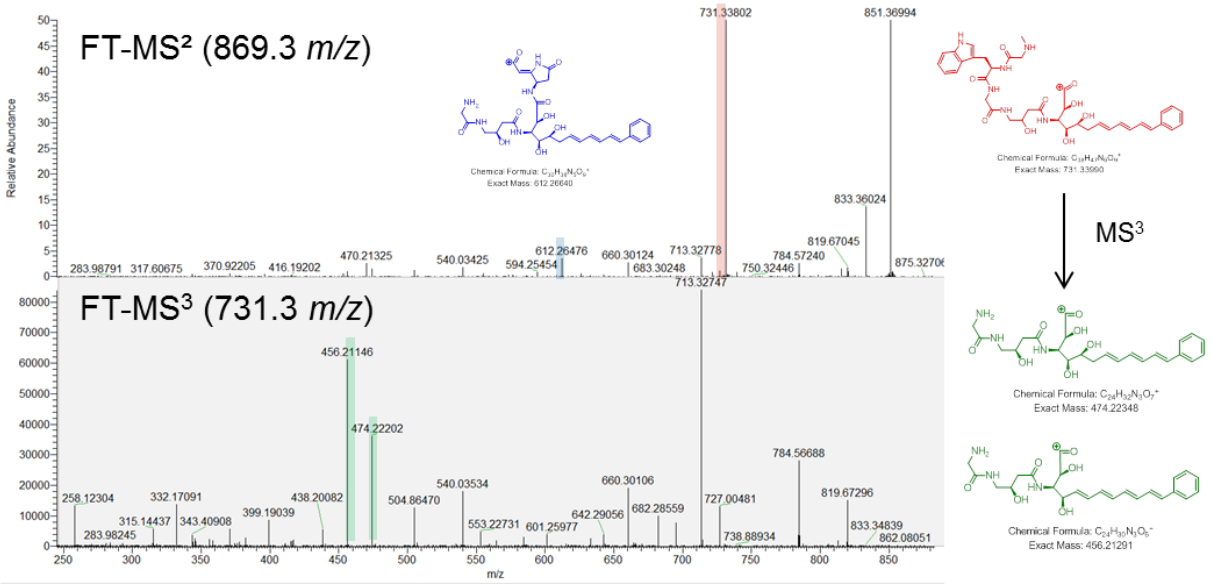


Figure S8: Accurate m/z measurement and CID fragmentation pattern of microsclerodermin M.

3.6.2.4 Pedein A

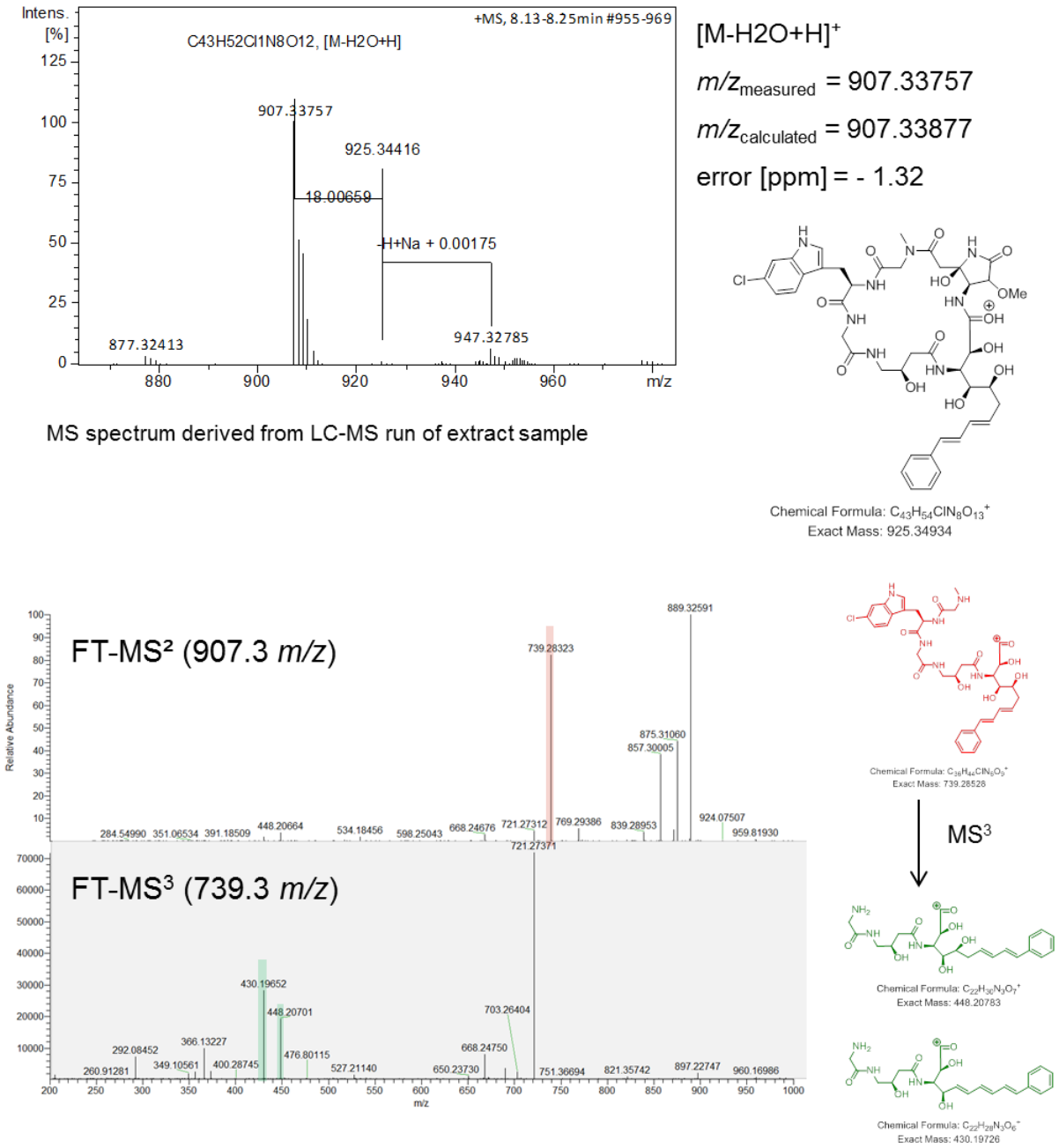


Figure S8: Accurate m/z measurement and CID fragmentation pattern of pedein A.

3.6.2.5 Pedein B

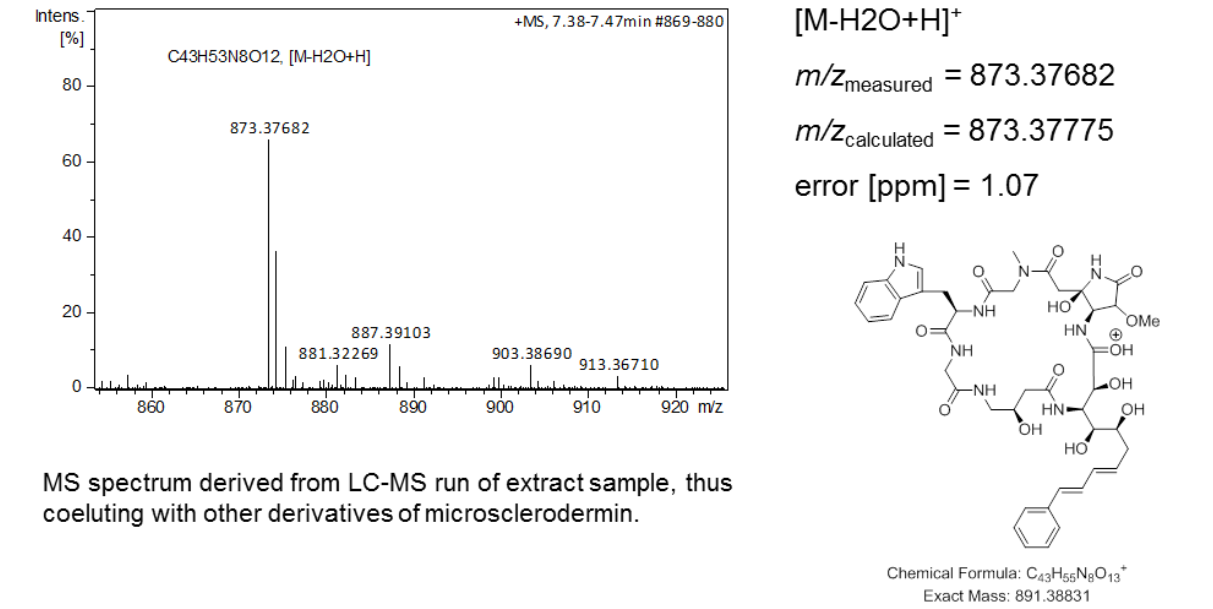


Figure S9: Accurate m/z measurement and CID fragmentation pattern of pedein B.

3.6.3 Analysis of the *msc* gene clusters in So ce38 and MSr9139

The regions of interest as well as 20 kb flanking regions were searched for open reading frames (ORFs) using Glimmer 3.0 and subsequently subjected to an automatic annotation by the antiSMASH software tool[1,2]. The *msc*-locus in So ce38 (Table S6) spans 58,049 bp and covers nine modules within the genes *mscA* to *mscI* with a GC content of 74.7 % (EMBL accession no. XXXX). The 5'-end of the gene cluster covers a major facilitator superfamily transporter (MFS-type, *mscK*) and a type II-thioesterase (*mscJ*) located approximately 2.5 kb upstream to *msc*A.

In the case of *Jahnella* sp. MSr9139 the cluster has a size of 62,268 bp and is thereby approximately 4.2 kb larger. We find the same genes as in So ce38 starting from *mscA* to *mscI* (Table S6, EMBL accession no. XXXX). There are three additional auxiliary genes named *mscL*, *mscM* and *mscN*. The first one encodes a halogenase specific for tryptophan found at the end of the cluster. This is consistent with the observation that a chlorinated tryptophan residue is found for the microsclerodermins isolated from *Jahnella* specimen. The genes *mscM* and *mscN* are located within the cluster (between *mscD* and *mscE*). They are encoding two proteins with significant similarity to an α-ketoglutarate dependent dioxygenase and a methyltransferase, respectively. The pairwise similarity of both clusters is 74.2 % on DNA-level. Both *msc* clusters have a comparable GC content of around 72 %.

Table S6: Genes involved in microsclerodermin biosynthesis as annotated in the strains Sorangium cellulosum So ce38 and Jahnella sp. MSr9139

	Sorangium cellulosum So ce38				Jahnella sp. MSr9139				
Gene	Start	End	Length [bp]	Function[1]	Start	End	Length [bp]	Function[1]	pairwise identity
mscA	3859	13686	9828	CoA-Lig-KR*-ACP-KS-AT-DH-KR-ACP	4124	14731	10608	CoA-Lig-KR* -MT-ACP-KS-AT-DH-KR-ACP	74.7
mscB	13709	16321	2613	KS-AT*	14718	17354	2637	KS-AT*	79.7
mscC	16327	20982	4656	KS-AT-DH-KR-ACP	17341	21990	4650	KS-AT-DH-KR-ACP	81.5
mscD	20995	23697	2703	KS-AT*	22000	24546	2547	KS-AT*	78.7
mscE	23727	25067	1341	Putative amidohydrolase	27907	26747	1161	Putative amidohydrolase	83.9
mscF	25335	32156	6822	PCP-ATT-MOX-C-A-PCP	28287	34856	6570	PCP-AAT-MOX-C-A-PCP	82.4
mscG	32158	36804	4647	KS-AT-KR-ACP	34858	39393	4536	KS-AT-KR-ACP	80.9
mscH	36844	49269	12426	C-A-M-PCP-C-A-PCP-E-C-A-PCP	39472	51792	12321	C-A-M-PCP-C-A-PCP-E-C-A-PCP	84.5
mscI	49344	58049	8706	C-A-PCP-KS-AT-KR-ACP-TE	51789	60668	8880	C-A-PCP-KS-At-KR-ACP-TE	83.9
mscJ	1473	2246	774	Thioesterase typeII	3006	3794	789	Thioesterase typeII	63
mscK	1248	1	1248	MFS-transporter	1350	1	1350	MFS-transporter	55.9
mscL	-	-	-	-	62268	60661	1608	Trp-halogenase	-
mscM	-	-	-	-	24583	25899	1317	Fe(II)/α-ketoglutarate dependent oxygenase	-
mscN	-	-	-	-	25881	26714	834	Methyltransferase	-

[1] A: adenylation domain, AAT: aminotransferase, ACP: acyl-carrier-protein domain AT: acyltransferase domain, C: condensation domain, CoA-Lig: CoA-Ligase, DH: dehydratation domain, E: epimerase, KR: ketoreductase domain, KS: ketosynthase domain, MT: methyltransferase domain, MOX: monooxygenase, PCP: peptidyl-carrier-protein domain, TE: thioesterase domain, *: inactive domain

Table S7: Proteins involved in microsclerodermin biosynthesis in the strains Sorangium cellulosum So ce38 and Jahnella sp. MSr9139.

	Sorangium cellulosum So ce38		Jahnella sp. MSr9139		
Protein	Length [aa]	Domains and position in sequence[a]	Length [aa]	Domains and position in sequence[a]	Identity [%]
MscA	3275	CoA-Lig (264-701), KR*(1034-1197), ACP (1348-1411), KS (1441-1828), AT (1976-2286), DH (2342-2505), KR' (2870-3047), ACP' (3149-3214)	3535	CoA-Lig (215-649), KR*(1032-1126), MT (1229-1509), ACP (1613-1676), KS (1712-2147), AT (2244-2555), DH (2610-2774), KR*' (3132-3309), ACP' (3411-3476)	65.4
MscB	870	KS (27-451), AT* (548-772)	878	KS (39-464), AT* (561-792)	73.4
MscC	1551	KS (36-460), AT (557-859), DH* (956-1076), KR (1158-1336), ACP (1436-1505)	1549	KS (39-464), AT (561-865), DH* (956-1076), KR (1157-1335), ACP (1436-1499)	75.0
MscD	900	KS (36-461), AT* (565-678)	848	KS (36-461), AT* (564-675)	72.0
MscE	446	Putative amidohydrolase	386	Putative amidohydrolase	82.9
MscF	2273	PCP (27-98), AMT (329-660), MOX (828-1127), C (1185-1529), A (1673-2082), PCP' (2169-2237)	2189	PCP (4-75), AMT (280-614), MOX (758-1057), C (1105-1397), A (1594-2000), PCP' (2088-2149)	76.5
MscG	1548	KS (14-439), AT (534-828), KR (1156-1330), ACP (1441-1509)	1511	KS (14-439), AT (531-850), KR (1136-1312), ACP (1404-1472)	72.1
MscH	4141	C (76-377), A (564-966), MT (1037-1256), PCP (1469-1531), C' (1554-1850), A' (2037-2426), PCP' (2515-2574), E (2591-2905), C'' (3074-3374), A'' (3558-3967), PCP'' (4054-4121)	4106	C (48-346), A (534-936), MT (1007-1225), PCP (1442-1505), C' (1526-1827), A' (2013-2405), PCP' (2492-2551), E (2568-2872), C'' (3043-3343), A'' (3527-3932), PCP'' (4019-4083)	78.8
MscI	2904	C (48-346), A (533-936), PCP (1023-1087), KS (1111-1535), AT (1638-1936), KR (2266-2465), ACP (2549-2612), TE (2634-2888)	2945	C (77-375), A (563-965), PCP (1053-1116), KS (1150-1573), AT (1676-1970), KR (2309-2509), ACP (2597-2660), TE (2683-2945)	77.8
MscJ	257	Thioesterase type II	263	Thioesterase type II	43.5
MscK	415	Major Facilitator Superfamily (MFS) transporter	450	Major Facilitator Superfamily (MFS) transporter	24.5
MscL	-	-	535	Trytophan halogenase	-
MscM	-	-	438	Fe(II)/α-ketoglutarate dependent oxygenase	-
MscN	-	-	277	Methyltransferase	-

[a] A: adenylation domain, AMT: aminotransferase, ACP: acyl-carrier-protein domain AT: acyltransferase domain, C: condensation domain, CoA-Lig: coenzyme A Ligase, DH: dehydratation domain, E: epimerase, KR: ketoreductase domain, KS: ketosynthase domain, MT: methyltransferase domain, MOX: monooxygenase, PCP: peptidyl-carrier-protein domain, TE: thioesterase domain, *: inactive domain

3.6.3.1 AT domains

Substrate specificity of acyl transferase (AT) domains was determined using conserved motif analysis by aligning the AT domains from both clusters to a reference AT from *E.coli* FAS, 1MLA (PDB 1MLA, UniProtKB P0AAI9) (Table S8).[3] This data shows that most of the listed AT-domains contain the GxSxG consensus motif with the catalytic residue in the center, except MscB in which the catalytic serine residue is replaced by a glycine. The MscD-AT domains are considered inactive as they are heavily truncated and do not align to the canonical motifs of AT domains. Active site analysis was used to predict which extender unit is recruited by a distinct AT domain. The main discriminant between using a methylmalonyl-CoA (mm-CoA) or malonyl-CoA (m-CoA) is the amino acid position 200 in 1MLA, where Ser200 supports mm-CoA and Phe200 supports m-CoA. The *in silico* prediction for microsclerodermin production is in all cases m-CoA and thereby in agreement with the chemical structure observed. We were not able to predict the incorporation of a hydroxymalonyl-CoA substrate for MscC-AT.

Table S8: Active site analysis of the acyl transferase domains (AT) from S. cellulosum So ce38 and Jahnella sp. MSr9139 according to Yadav et al.[3]

Domain	observed	predicted	11	63	90	91	92	93	94	117	200	201	231	250	255	15	58	59	60	61	62	70	72	197	198	199
1MLA	M	M	Q	Q	G	H	S	L	G	R	S	H	N	Q	V	T	K	T	W	Q	T	S	A	S	V	P
MscA-AT$_{So\ ce38}$	M	M	Q	Q	G	H	S	V	G	R	F	H	N	H	V	Q	R	T	E	Y	T	E	A	S	H	A
MscA-AT$_{Jahnella\ sp.}$	M	M	Q	Q	G	H	S	V	G	R	F	H	N	H	V	Q	R	T	E	Y	T	E	A	S	H	A
MscB-AT$_{So\ ce38}$	inactive	-	E	Q	G	Q	G	A	G	R	-	-	S	D	D	T	Q	T	A	F	T	Q	A	-	-	-
MscB-AT$_{Jahnella\ sp.}$	inactive	-	Q	Q	G	L	G	V	G	R	-	-	S	Q	D	L	Q	A	A	F	A	E	A	-	-	-
MscC-AT$_{So\ ce38}$	Hydroxy malonate	M	Q	Q	G	H	S	I	G	R	F	H	N	H	V	Q	Q	T	A	F	T	E	A	S	H	A
MscC-AT$_{Jahnella\ sp.}$	Hydroxy malonate	M	Q	Q	G	H	S	I	G	R	F	H	N	H	V	F	Q	T	A	F	A	E	A	S	H	A
MscG-AT$_{So\ ce38}$	M	M	Q	Q	G	H	S	I	G	R	F	H	N	H	V	H	D	T	A	I	A	E	A	S	H	A
MscG-AT$_{Jahnella\ sp}$	M	M	Q	Q	G	H	S	V	G	R	F	H	N	H	V	H	D	T	A	L	A	G	A	S	H	A
MscI-AT$_{So\ ce38}$	M	M	Q	Q	G	H	S	V	G	R	F	H	N	H	V	Y	Q	T	R	L	T	E	A	S	H	A
MscI-AT$_{Jahnella\ sp}$	M	M	Q	Q	G	H	S	V	G	R	F	H	N	H	V	H	Q	T	R	L	A	E	A	S	H	A

3.6.3.2 KS domains

The ketosynthase (KS) domains of both msc biosynthetic gene clusters were extracted and analyzed using the NaPDoS web tool.[4] The result is depicted as a phylogenetic tree in Figure S11 and allows easy annotation of the KS subtypes found in the msc cluster.

The postulated iterative function of MscB is supported by grouping of these domains at the interface between modular (ochre) and iterative (green) domains. The remaining, functional domains MscA, MscG, and MscI are correctly grouped in the PKS/NRPS hybrid section. MscA is characterized as a hybrid domain because of its linkage to the first module with its uncommon starter unit.

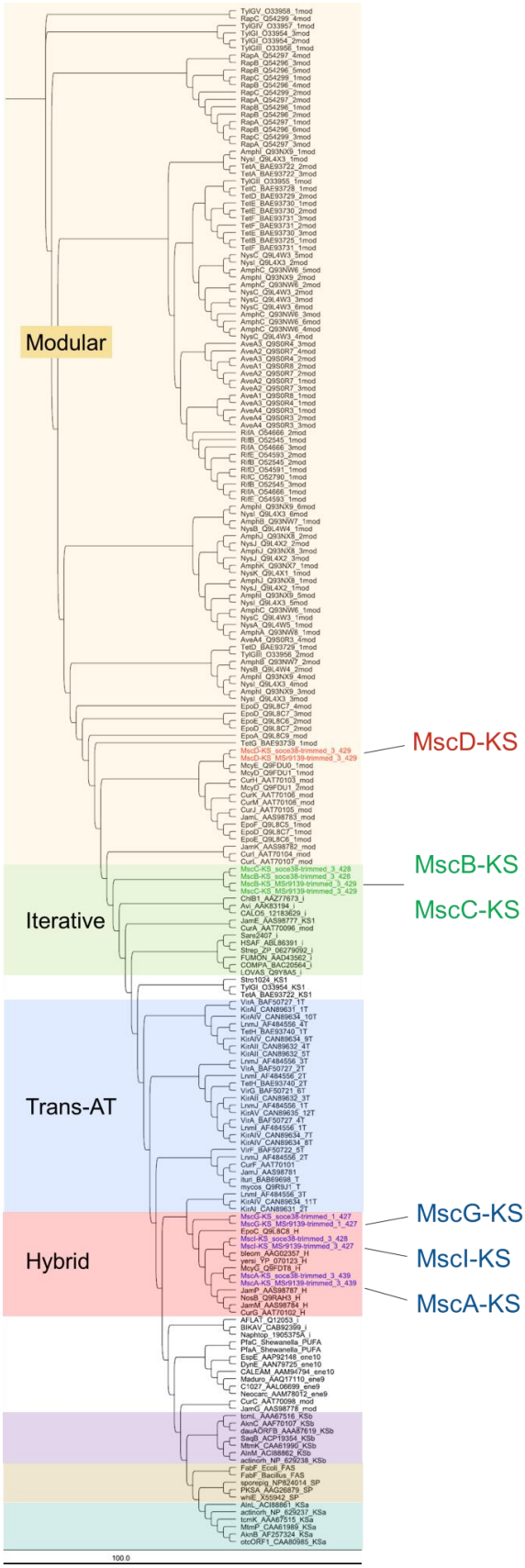


Figure S11: Phylogenetic tree highlighting different KS domain types and the KS domains found in the msc biosynthetic gene cluster in So ce38 and MSr9139.

3.6.3.3 DH domains

Both *msc* clusters contain two dehydration domains present in the proteins MscA and MscC. Based on the proposed biosynthesis of microsclerodermin the DH-domain in MscA must be active. However, it does not exhibit the common consensus-motif $H(X_3)G(X_4)P$ of DH-domains as known from other modular PKS-systems (Table S9). Here, the central glycine is replaced by a valine. This substitution is obviously not affecting the enzyme function. The DH-domains from MscC are considered to be inactive as they do not contain a conserved motif at all. The domains are truncated with a protein length of around 100 amino acids.

Table S9: Active site analysis of the dehydratase domains (DH) of the msc biosynthetic gene cluster.

Consensus	H	x	x	x	G	x	x	x	x	P
MscA-DHSo ce38	H	A	E	G	V	L	V	A	R	P
MscA-DHJahnella sp	H	V	E	G	V	V	V	A	R	P

3.6.3.4 KR domains

The *msc* cluster from both myxobacteria encodes five ketoreductase domains. The KR domains may be classified as A- and B-type by analyzing the amino acid sequence of two regions within the amino acid sequence.[5] However, it was shown that a reliable prediction of the stereogenic center that emerges upon KR reduction of a keto-function is governed by additional, unknown interactions.[6] Nevertheless, we attempted to classify the domains in this work. The occurrence of a LDD motif is a good indicator for a B-type KR domain. This is sometimes supported by the presence of a Pro144 and an Asp148 according to the numbering based on the DEBS2 KR domain (GenBank #X62569.1). In case of A-type KR domains a highly conserved Trp141 is present. Moreover, a correct co-factor binding motif is mandatory for the enzyme function. Based on this, MscA-KR1 of both strains is predicted to be inactive as it does not exhibit the co-factor binding motif GxGxxGxxxA (Table S10). Additionally, the MscA-KR1 domain of MSr9139 is heavily truncated. The remaining KR domains feature the correct binding motif.

Table S10: Analysis of the co-factor binding motif of the ketoreductase domains (KR) within the msc biosynthetic gene cluster.

Domain	G	x	G	x	x	G	x	x	x	A	motif
MscA-KR1 $_{\text{So ce38}}$	D	L	G	G	V	S	L	Q	L	L	inactive
MscA-KR1 $_{\text{MSr9139}}$	A	L	G	S	Y	E	-	-	-	-	inactive
MscA-KR2 $_{\text{So ce38}}$	G	L	G	A	L	G	R	R	V	A	ok
MscA-KR2 $_{\text{MSr9139}}$	G	L	G	A	L	G	R	R	V	A	ok
MscC-KR$_{\text{So ce38}}$	G	T	G	A	L	G	L	H	I	A	ok
MscC-KR$_{\text{MSr9139}}$	G	I	G	A	L	G	L	H	T	A	ok

MscG-KR$_{So\ ce38}$	G	L	G	G	A	G	L	G	I	A	ok
MscG-KR$_{MSr9139}$	G	L	G	G	V	G	L	A	I	A	ok
MscI-KR$_{So\ ce38}$	G	L	G	R	I	G	L	C	L	A	ok
MscI-KR$_{MSr9139}$	G	L	G	R	I	G	L	C	L	A	ok

The two amino acid regions for KR domain classification are shown in Table S11. MscA-KR2 and MscC-KR1 can be classified as B-type and A-type, respectively. The LDD-motif together with an Asp148 is found in MscA-KR2 whereas the highly conserved Trp141 specifies MscC-KR. One stereogenic center originating from the A-type MscC-KR is still present in the final product and should exhibit *S*-configuration. This is in agreement with the configuration of the last hydroxyl group reported for the microsclerodermin side chain. The KR domains of MscG do only feature a LDD-like motif but no additional identifier in the second region. It remains unclear whether this domain is active at all. MscI-KR produces an *R*-configured hydroxyl moiety as identified by advanced Marfey analysis of microsclerodermin (GABA subunit). However, there is no common consensus that allows a classification. Moreover, the KR domains exceed the common KR domain length of approximately 180 amino acids by additional 30 amino acids.

Table S11: Analysis of the ketoreductase domains (KR) within the msc biosynthetic gene cluster

	region 88-103																region 134-149														KR type
MscA-KR2 $_{So\ ce38}$	H	L	A	G	A	L	D	D	G	V	L	L	Q	Q	S	W	F	S	S	I	A	S	V	L	G	S	A	A	Q	G	N	Y	B-type
MscA-KR2 $_{MSr9139}$	H	L	A	G	S	L	D	D	G	V	L	V	Q	Q	S	W	F	S	S	I	A	S	I	L	G	S	A	G	Q	G	N	Y	B-type
MscC-KR$_{So\ ce38}$	H	A	A	G	V	A	G	A	R	D	L	S	A	L	D	A	T	S	S	I	A	S	L	W	G	S	R	G	Q	A	H	Y	A-type
MscC-KR$_{MSr9139}$	H	A	A	G	V	G	G	Y	C	E	L	S	R	L	D	A	Y	S	S	I	A	S	L	W	G	S	R	G	Q	A	H	Y	A-type
MscG-KR$_{So\ ce38}$	Y	A	A	G	V	D	D	P	A	L	I	G	-	G	I	G	S	S	S	L	S	A	I	L	G	G	R	G	L	G	A	Y	B-type
MscG-KR$_{MSr9139}$	Y	A	A	A	I	D	D	A	G	L	L	G	-	A	V	G	G	S	S	L	S	A	I	L	G	G	R	G	L	A	A	Y	B-type
MscI-KR$_{So\ ce38}$	H	A	A	G	A	R	G	D	G	T	F	M	V	P	L	A	L	S	S	T	S	A	I	L	G	G	L	G	L	G	P	Y	unknown
MscI-KR$_{MSr9139}$	H	A	A	G	A	R	G	D	G	T	F	M	T	S	L	A	L	S	S	T	S	A	I	L	G	G	L	G	L	G	P	Y	unknown

3.6.3.5 A domains

Three genes of microsclerodermin biosynthesis (*mscF*, *mscH* and *mscI*) harbor at least one NRPS module. In order to determine the A domain specificity of each module the respective amino acid sequences were analyzed using the NRPSpredictor2 tool.[7] The results are listed in Table S12. The predicted specificity of each A-domain is consistent with the observed amino acid incorporated into the microsclerodermin scaffold.

Table S12: Prediction of A-domain specificity for each A-domain present in the msc biosynthetic gene cluster using the NRPSpredictor2 software-tool.[7]

	amino-acid specificity	
domain	So ce38	MSr9139
MscF-A	asn	asn
MscH-A1	gly	gly
MscH-A2	trp	trp
MscH-A3	gly	gly
MscI-A	gly	gly

3.6.3.6 C domains

All C domains of the *msc* biosynthetic gene cluster were analyzed using the NaPDoS web tool.[4] Especially the hybrid module MscF is of interest as the incorporated *S*-asparagine is (a) cyclized to form and pyrrolidone moiety and (b) converted to the *R*-asparagine. The mechanism behind these reactions is not understood yet. Analysis of the respective C domain, MscF-C1, characterizes the domain as a common PKS/NRPS hybrid domain (Figure S12). Hence, the observed cyclization as well as inversion of asparagine is not done by the C domain. It is more likely related to the amidohydrolase MscE which is found in both clusters. Its detailed function remains elusive at the current stage of research.

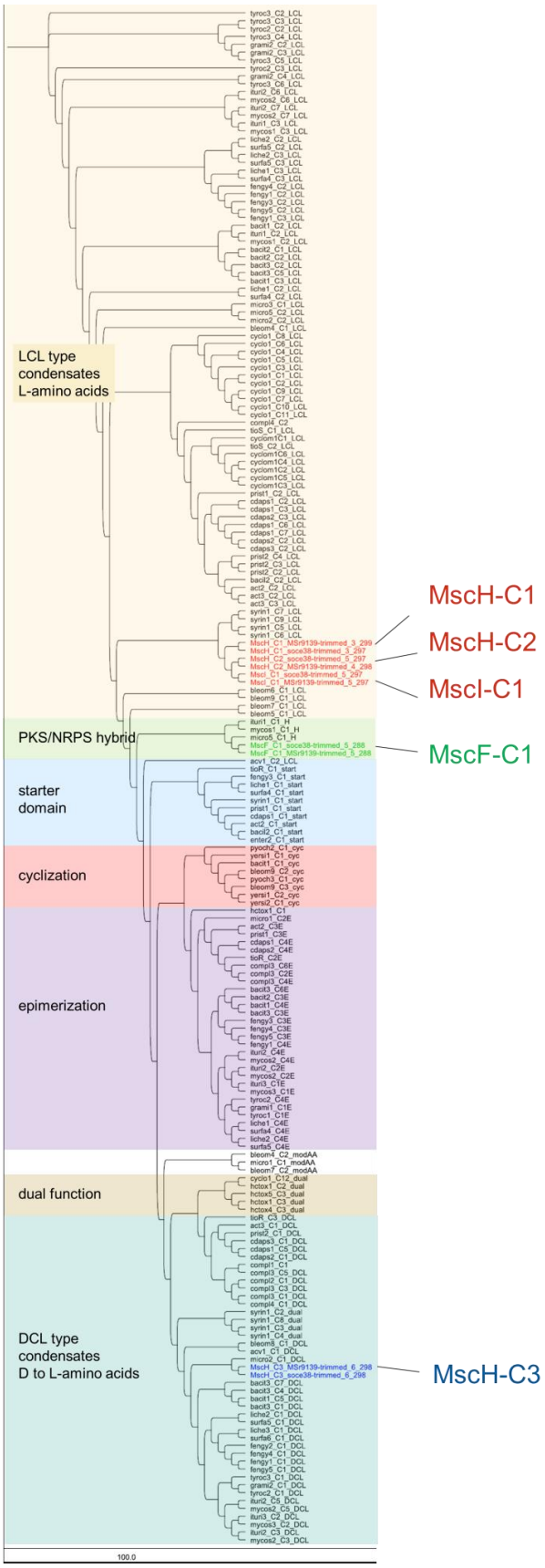


Figure S12: Phylogenetic tree highlighting different C domain types and the C-domains found in the msc biosynthetic gene cluster in So ce38 and MSr9139.

3.6.3.7 MT-domains

The *N*-methyltransferase present in module 1 of MscH is found in both clusters and is responsible for the *N*-methylation of the glycine right next to the pyrrolidone ring. Both of them harbor the common consensus-motif GxGxG at the *N*-terminal end which is responsible for SAM Co-factor binding (Table S13).[8] These *N*-methyltransferases belong to the UbiE/COQ5 family. In addition, the *msc*-cluster of *Jahnella* sp. encodes a second methyltransferase domain which is located in MscA. This methyl transferase has homology to the PRMT5 family. However, the MT domain is likely inactive as the SAM binding motif GxGxG is changed to GxGxE. The exchange of the small glycine to a charged glutamic acid may prevent cofactor binding.

Table S13: Active site analysis of the methyltransferase domains (MT) within the msc biosynthetic gene cluster.

active consensus	G	x	G	X	G
MscA-MTJahnella sp.	G	T	G	K	E
MscH-MTSo ce38	G	C	G	T	G
MscH-MTJahnella sp.	G	C	G	T	G

3.6.4 Determination of configuration

The conformation of all stereogenic centers of the isolated microsclerodermins is identical to that of the microsclerodermins C – I and confirms the assumed stereochemistry of the pedeins.[9–11] In detail, the scaffold is based on an *R*-tryptophan, a (*3R*)-4-NH_2-3-OH-butyric acid moiety, a side chain with (*2S,3R,4S,5S*)-configuration and an (*R,R*)-pyrrolidone.

3.6.4.1 Absolute configuration of 3-aminobutyric acid and tryptophan

Microsclerodermin has several stereogenic centers. We identified the configuration of the γ-aminobutyric acid (GABA) moiety and tryptophan by acid hydrolysis followed by advanced Marfey analysis using FDLA reagent. (3*S*)-GABA was ordered from Sigma and used as a reference together with *S*-tryptophan and *S*-6-chloro-tryptophan. Derivatization of the γ-amino functionality with L-FDLA and D-FDLA results in two products with reproducible retention time differences (3 seconds using 10 technical replicates, see Figure S13) which allows annotation of the stereochemistry. In conclusion, the new microsclerodermins L and M have an (3*R)*-configured carbon atom in the GABA subunit which is in agreement with the other microsclerodermins reported.

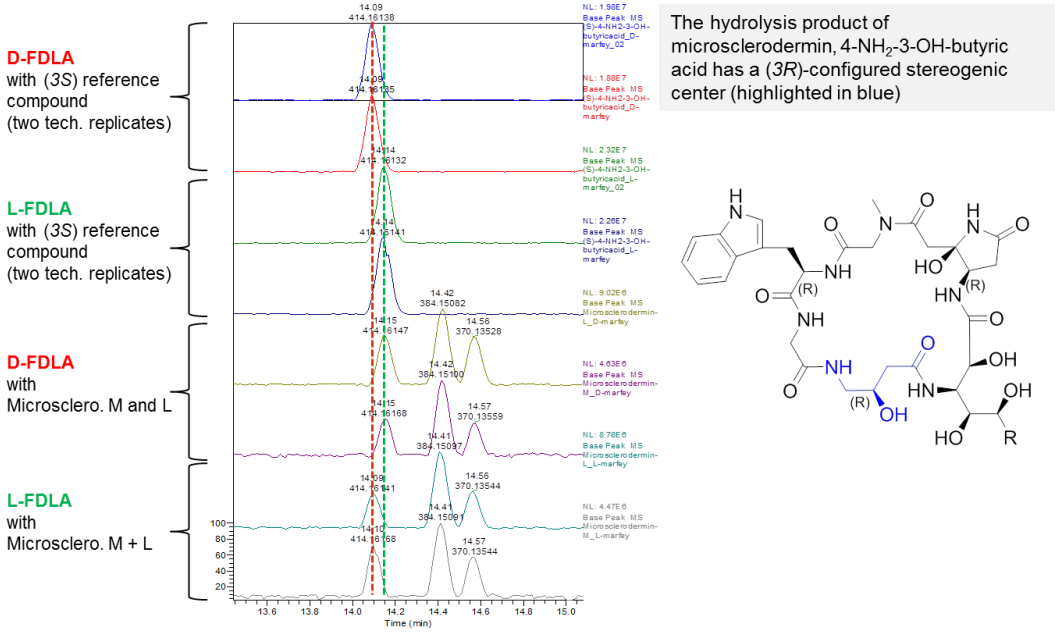


Figure S13: Overlay of LC-MS chromatograms highlighting the retention time difference of the GABA subunit when derivatized with Marfey's reagent (FDLA). The graph features the (3S)-GABA reference compound as well as microsclerodermin M and L.

The tryptophan in microsclerodermin M, D and L is *R*-configured. This applies to all microsclerodermins known so far and is in agreement with the epimerization domain found within the respective module of the biosynthetic assembly lines in So ce38 and MSr9139. Microsclerodermin M was compared to *S*-tryptophan whereas microsclerodermin D and L were compared to a *S*-6-Cl-tryptophan reference (Figure S14).

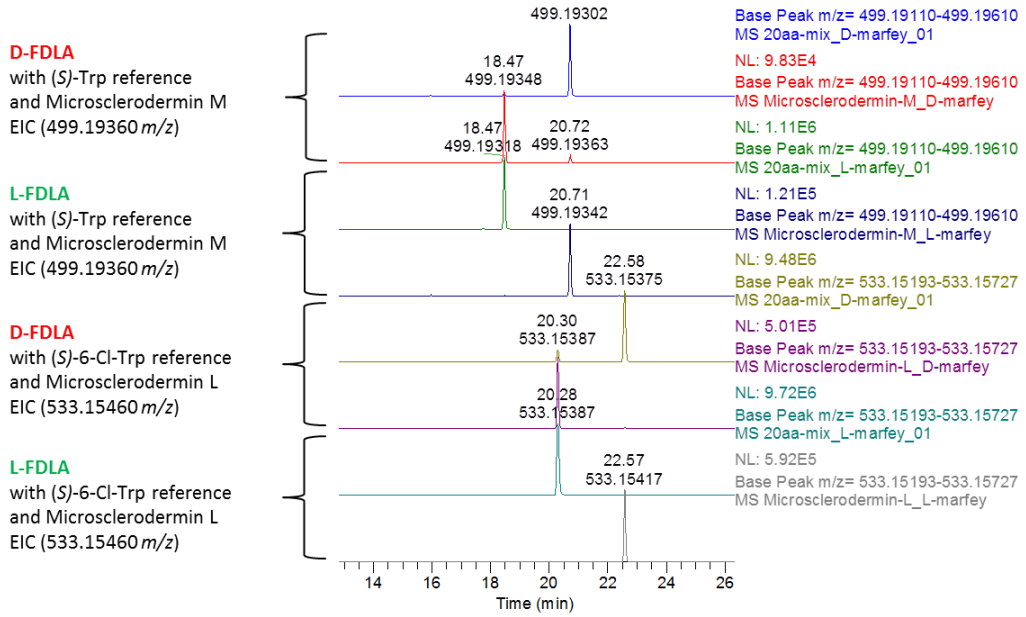


Figure S14: Overlay of the LC-MS chromatograms for tryptophan characterization using advanced Marfey's reagent (FDLA). The stereogenic center of tryptophan is (R)-configured in microsclerodermin M and L.

3.6.4.2 Configuration of the pyrrolidone moiety

The pyrrolidone moiety of the microsclerodermins is derived from an asparagine as indicated by feeding labeled *S*-asparagine. Incorporation of labeled *S*-asparagine suggested that the pyrrolidone ring keeps this configuration. However, this turned out to be wrong. Microsclerodermins M, D and L underwent the same chemical degradation steps as reported for the microsclerodermins A-I.[9,10,12] After dehydration at the pyrrolidone moiety the respective dehydromicrosclerodermins were subjected to ozonolysis. The reaction products were hydrolyzed and derivatized with Marfey's reagent. Marfey derivatization and HPLC-MS measurements revealed an *R*-configured pyrrolidone moiety in the microsclerodermin derivatives M, L, and D. Although this chemical degradation protocol is used for the microsclerodermins A – I in literature, the method is not that elegant in terms of the result as ozonolysis, oxidative workup and hydrolysis of *R*-tryptophan will result in *R*-asparate as well. At the same time this procedure will generate aspartate out of the pyrrolidone moiety. By doing a short hydrolysis step of not more than 60 minutes we observe a notable amount of *R*-asparagine in the sample which can only originate from the pyrrolidone moiety. Based on these results we were able to identify *R*-tryptophan and *R*-pyrrolidone in microsclerodermin M and D. We verified the results by applying the above mentioned protocol to pure *R*-tryptophan as well as to the synthesized, enantiomeric pure *S*-dehydropyrrolidone fragment (98 % ee). *R*-tryptophan is converted to *R*-aspartate while the *S*-dehydropyrrolidone fragment is converted to asparagine and aspartate with a *S*:*R*-ratio of approximately 4:1. This partial inversion clarifies the origin of the *R*- and *S*-asparagine mixture which is observed in all samples of microsclerodermin, there with the opposite *S*:*R*-ratio of 1:4 (Figures S15)

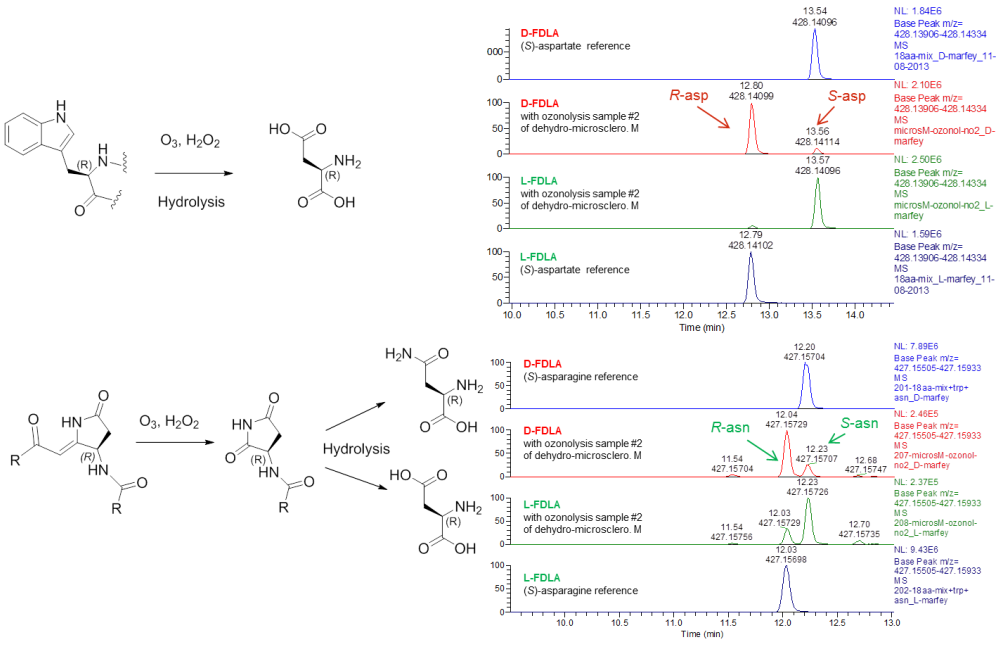


Figure S15: Marfey analysis result of ozonolysis product of microsclerodermin M. The chromatograms show the peaks for asparagine and aspartate, both as an overlay display with the respective S-configured reference compound.

Microsclerodermin D and L are both isolated from strain MSr9139. In case of microsclerodermin L, the pyrrolidone moiety has an additional methoxy group resulting in a 3-methoxy-aspartate after the above mentioned reactions whereas microsclerodermin D has the same pyrrolidone moiety as the M derivative (So ce38). Unfortunately we were not able to get suitable enantiomeric pure reference compounds for 3-methoxy-aspartate. However, based on the rule of thumb that an L-Marfey derivative of an L-configured amino acid elutes earlier than that of an D-Marfey derivative[13] we can conclude that it is the other way round in our sample and we do have an D-configured amino acid, (*2R*)-3-methoxy-aspartate, present in microsclerodermin L (Figure S16). This is supported by the detection of *R*-asparagine in the microsclerodermin D derived sample which is isolated from the same strain and thereby underlying the same biosynthesis. The stereogenic center at the methoxy group was not assigned by this method.

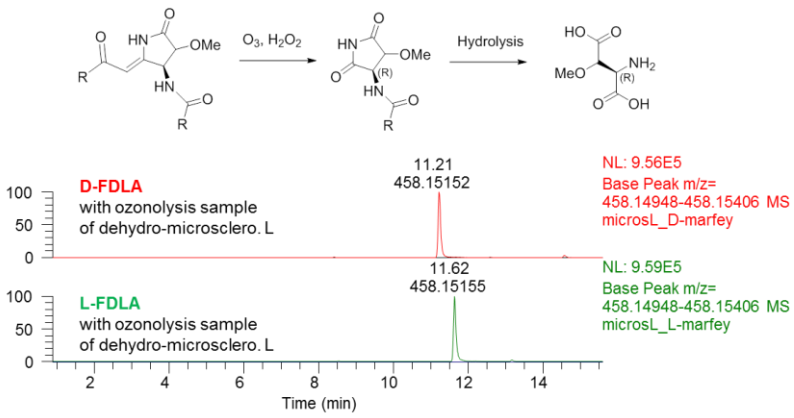


Figure S16: Marfey analysis result of ozonolysis product of microsclerodermin L. The chromatograms show the peaks for 3-MeO-aspartate. As the D-marfey derivative (upper part) elutes before the L-marfey derivative (lower part) the 3-MeO-aspartate has a 2R-configured stereogenic center. There is no reference compound available.

3.6.4.3 Configuration of the side chain

The side chain's vicinal hydroxyl groups OH-4 and OH-5 were converted to an acetonide in order to identify the relative stereochemistry of this part of the molecule. The acetonide formation was done for microsclerodermin M from *S. cellulosum* So ce38 and microsclerodermin L from *Jahnella* sp. MSr9139. Acetonides were purified using the same HPLC method as for the natural product isolation. Purity was checked by HPLC-MS analysis prior to NMR analysis. Respective NOE couplings are detected using selective irradiation experiments for the protons attached to the acetonide moiety. The NOE couplings are identical to those determined for the microsclerodermins known to literature. Based on this result we conclude to have the same relative stereochemistry as reported for microsclerodermins A – I which have a *2S,3R,4S,5S*-configured side chain. We assigned the absolute stereochemistry of the side chain by diol cleavage at OH-4 and OH-5. The product is oxidized to the respective carboxylic acid and hydrolyzed

using 6 N HCl. After Marfey derivatization we observed a peak for L-threo-ß-OH-aspartate (*2S,3S* configuration) which would be in agreement with the *2S,3R,4S,5S*-configured side chain.

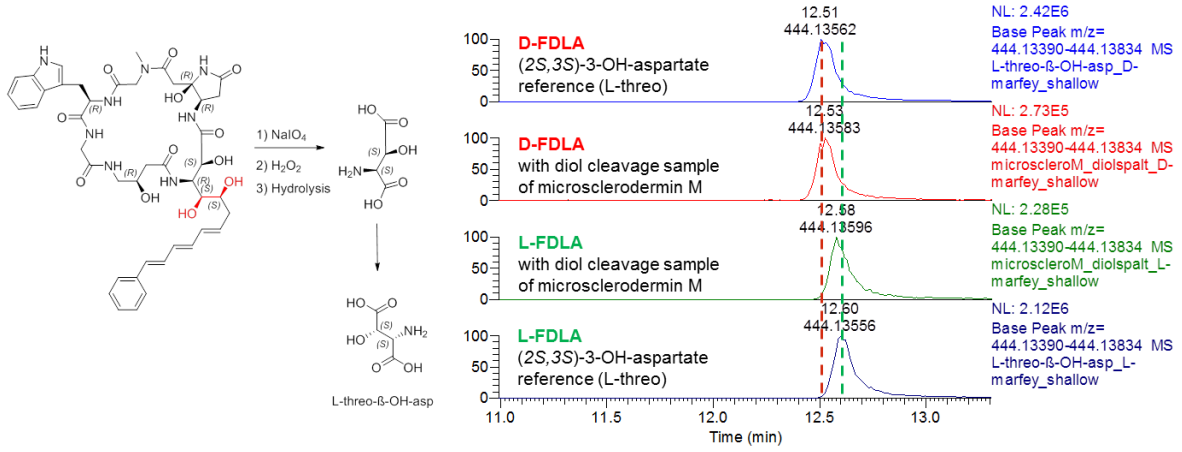


Figure S17: Marfey analysis of diol cleavage product (Microsclerodermin M).Peak identified by reference compound L-threo-ß-OH-aspartate.

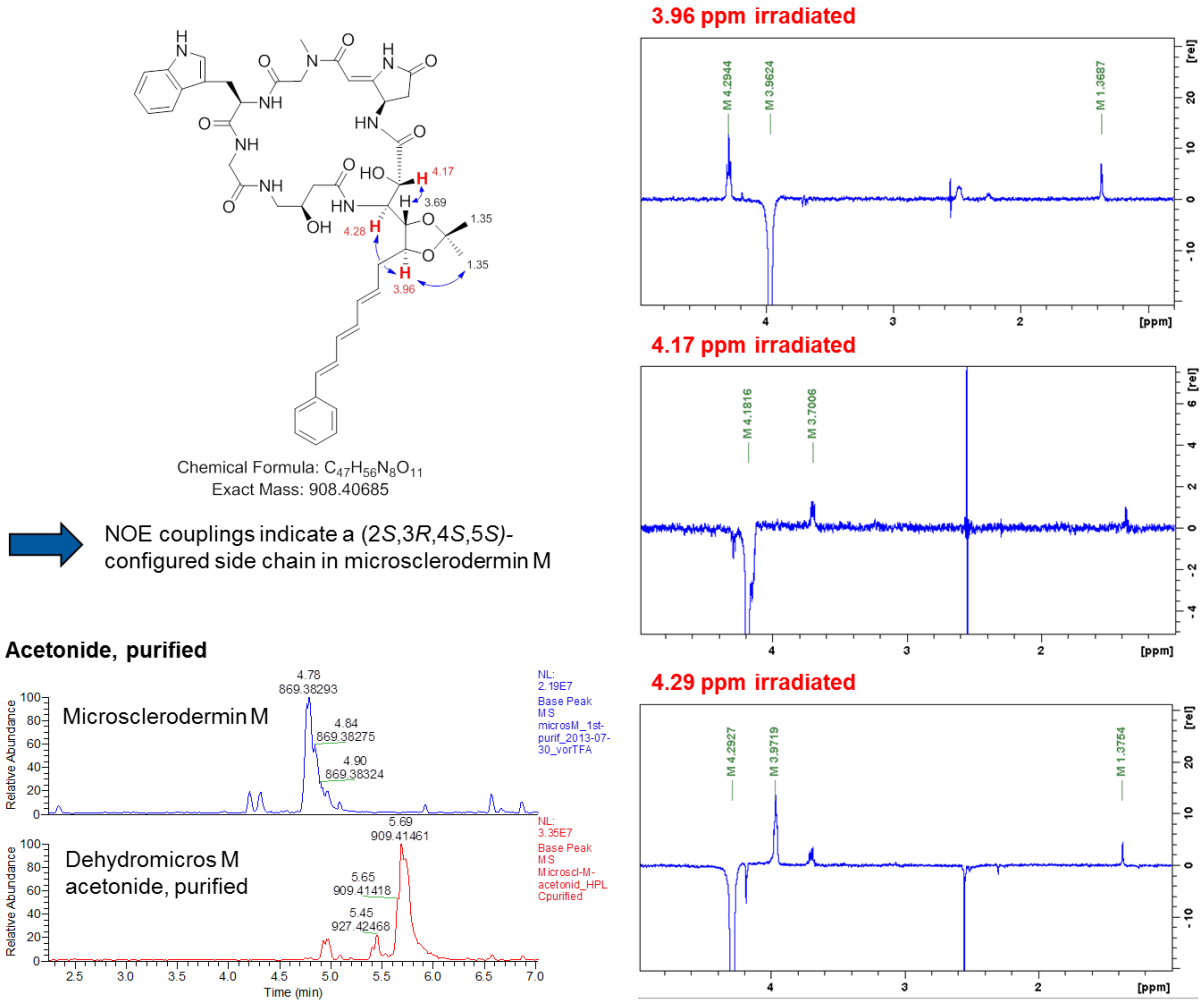


Figure S18: Acetonide purification and selective irradiation experiments of dehydromicrosclerodermin M.

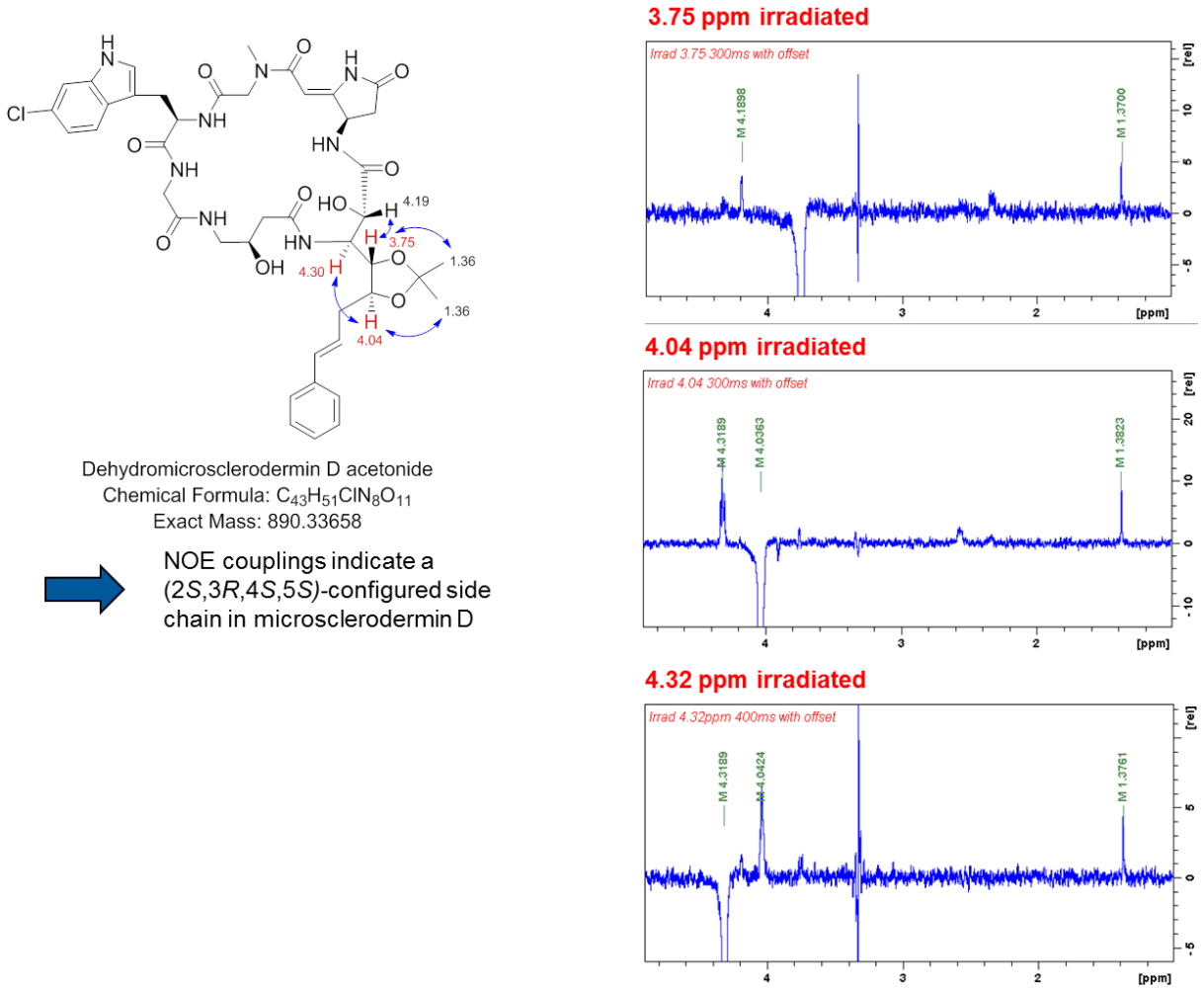


Figure S19: Acetonide purification and selective irradiation experiments of dehydromicrosclerodermin D.

3.6.5 Experimental section

3.6.5.1 Dehydration of microsclerodermin

A solution of microsclerodermin (2 mg) in 100 µl DMSO was diluted with 1 % aqueous TFA (400 µl). The mixture was kept at 40 °C for 15 minutes and dried afterwards using a GeneVac Evaporator (GeneVac Ltd, Suffolk, UK). Quantitative conversion to the respective dehydromicrosclerodermin was verified by dissolving the products in methanol and measuring them by LC-MS analysis.

3.6.5.2 Ozonolysis of dehydromicrosclerodermin

Dehydromicrosclerodermin (100 µg) was dissolved in 400 µL dry methanol. A stream of ozone was bubbled through the cooled solution (-78 °C) for 20 minutes. The excess reagent was removed by a stream of nitrogen before warming the solution to 20 °C. Remaining solvent was then removed under reduced pressure and the dry ozonide dissolved in 1 mL formic acid/hydrogen peroxide (2:1, v/v). The reaction mixture was heated to 80 °C for 30 min before drying the mixture under reduced pressure. The sample was then used for hydrolysis and Marfey derivatization.

3.6.5.3 Diol cleavage of microsclerodermin

Microsclerodermin (500 µg) was dissolved in 50 µL DMSO and added to 300 µL of sodium periodate solution in water (0.5 mg/mL), adjusted to pH 4.0 with acetic acid. The sample was stirred at room temperature for 16 hours and dried afterwards. The residue was dissolved in hydrogen peroxide (1 mL) and formic acid (0.5 mL) and heated to 70 °C for 20 min. After removing the solvent under reduced pressure the sample was ready for hydrolysis and Marfey derivatization.

3.6.5.4 Acetonide of dehydromicrosclerodermin

Under an atmosphere of dry nitrogen 200 µL of 2,2-dimethoxy propane is added to a stirred solution of dehydromicrosclerodermin (2 mg) in 150 µL dry DMF. The reaction is initiated by adding catalytic amounts of pyridinium p-toluene sulfonic acid and stirred for 16 hours at room temperature. The reaction is quenched by adding 20 µl pyridine and dried using a stream of nitrogen. The residues are diluted in 100 µL H_2O/DMSO (1:1, v/v) and the respective acetonide purified using the same HPLC method as for the initial purification of the microsclerodermins in this work.

3.6.5.5 Marfey derivatization protocol

Marfey derivatization was performed with NMR, ozonolysis and diol-cleavage samples of the compounds. The protocol is as follows:

- Put at least 50 µg of sample into a 1.4 mL glass vial
- Add 100 µL of 6 N HCl. Fill the vial with nitrogen and close it. Keep it at 110 °C for 45 minutes. Open the vial to let dry at 110 °C for another 15 minutes. Do not exceed 60 min as tryptophan easily decomposes.
- Dissolve the residues in 110 µL H_2O. Prepare two 1.5 mL PP tubes and add 50 µL of the aqueous solution in each.
- Add 20 µL of 1 N $NaHCO_3$ in each tube (pH adjusted to approx. 9)
- Add 20 µL of 1 % (w/v) Marfey's reagent in acetone: D-FDLA or L-FDLA, respectively.
- Keep for 1 hour at 40 °C, 700 rpm.
- Add 10 µL of 2 N HCl to stop reaction and 300 µL of ACN to end up with 400 µL total volume.
- Centrifuge the sample and measure the supernatant by HPLC-MS

3.6.5.6 HPLC-MS analysis of Marfey samples

All measurements were performed on a Dionex Ultimate 3000 RSLC system using a Waters BEH C18, 100 x 2.1 mm, 1.7 µm d_p column by injection of 1 µl sample. Separation was achieved by a gradient using (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 550 µl/min and 45 °C. The gradient was as follows: starting at 5 % B to increase to 10 % B in 1 min, from 1 to 15 min increase to 35 % B, from 15 to 22 min increase to 50 % B, from 22 to 25 min increase to 80 % B. After a 1 min hold at 80 % B the system was reequilibrated with initial conditions for 5 minutes. UV data was acquired at 340 ± 8 nm and MS detection was performed simultaneously. Coupling the HPLC to the MS was supported by an Advion

Triversa Nanomate nano-ESI system attached to a Thermo Fisher Orbitrap. LC flow is split to 500 nL/min before entering the ion source. Mass spectra were acquired in centroid mode ranging from 150 – 1000 m/z at a resolution of R = 30000.

3.6.6 HPLC-MS based screening

We searched our in-house database for producers of microsclerodermins. This screening included LC-MS data from 791 myxobacterial extracts covering a broad diversity of myxobacterial genera (Figure S20). Eleven different *Sorangium* sp. were identified as producers of microsclerodermin M. Microsclerodermin M is a rather instable compound that readily forms isobaric isomers. This effect is most likely related to the unsaturated side chain and causes several peaks that nearly coelute under standard LC-MS screening conditions. Thus, extracted ion chromatograms (EICs) for the $[M\text{-}H_2O+H]^+$ signal at 869.38283 *m/z* (blue trace, Figure S21) show up with a broadened peak including multiple isomers. In addition, the $[M+H]^+$ signal at 887.39340 *m/z* is detected albeit with lower abundance. We observe an additional, later eluting peak in the EICs with a $[M+H]^+$ signal at 869.38283 *m/z*, most likely caused by a loss of water at the pyrrolidone ring upon extract preparation and storage.

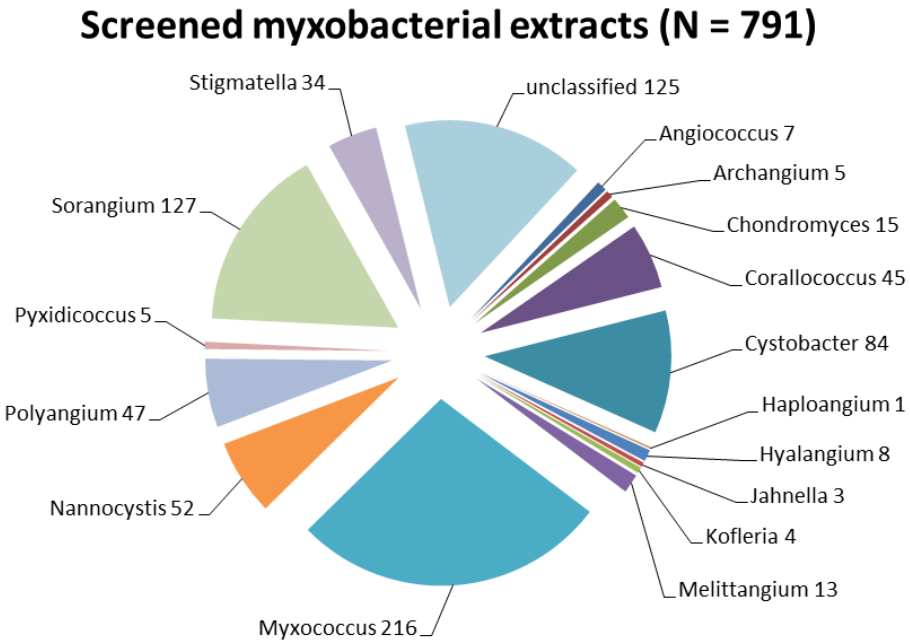


Figure S20: Myxobacterial genera as covered in our LC-MS based screening data set.

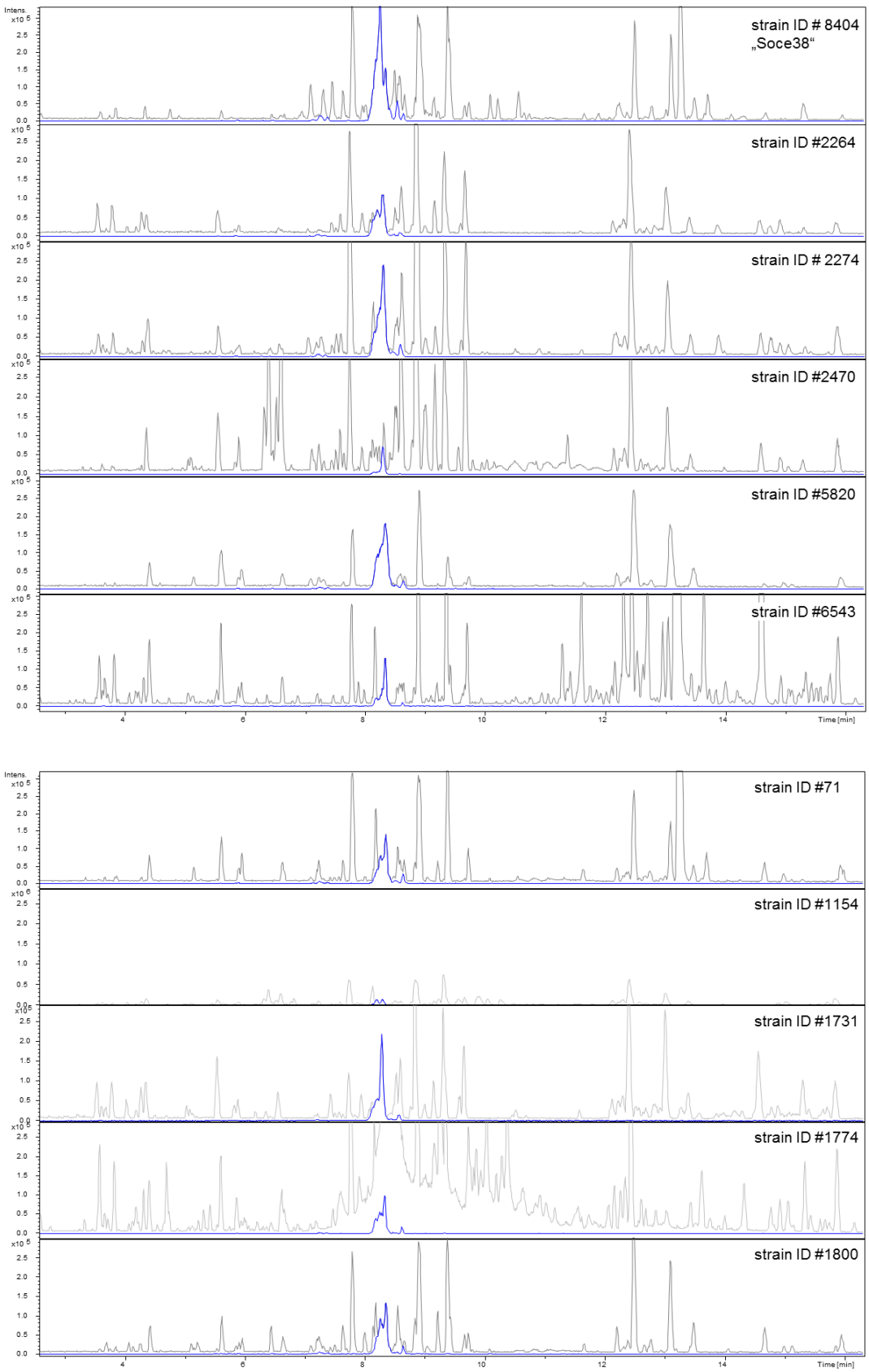


Figure S21: List of microsclerodermin M producers as identified by searching our in-house database. All strains are of the genus Sorangium. Microsclerodermin M and its isobaric isomers are highlighted by the EIC of 869.38283 m/z ($[M-H_2O+H]^+$ ion, blue trace). Abundance of derivatives is related to the different producer strain. Irregular peak shape is due to isobaric, almost coeluting derivatives occurring during storage of the extracts. Identification of microsclerodermin M is based on accurate m/z, isotope pattern fit, retention time and MS^2 pattern compared to the reference compound.

Another group of myxobacteria is producing the known microsclerodermin D as well as pedein A and B. In addition, one new derivative (Microsclerodermin L) is identified in this work (Figure S22). The strain *Chondromyces pediculatus* Cm p3 is also part of this group.[11]

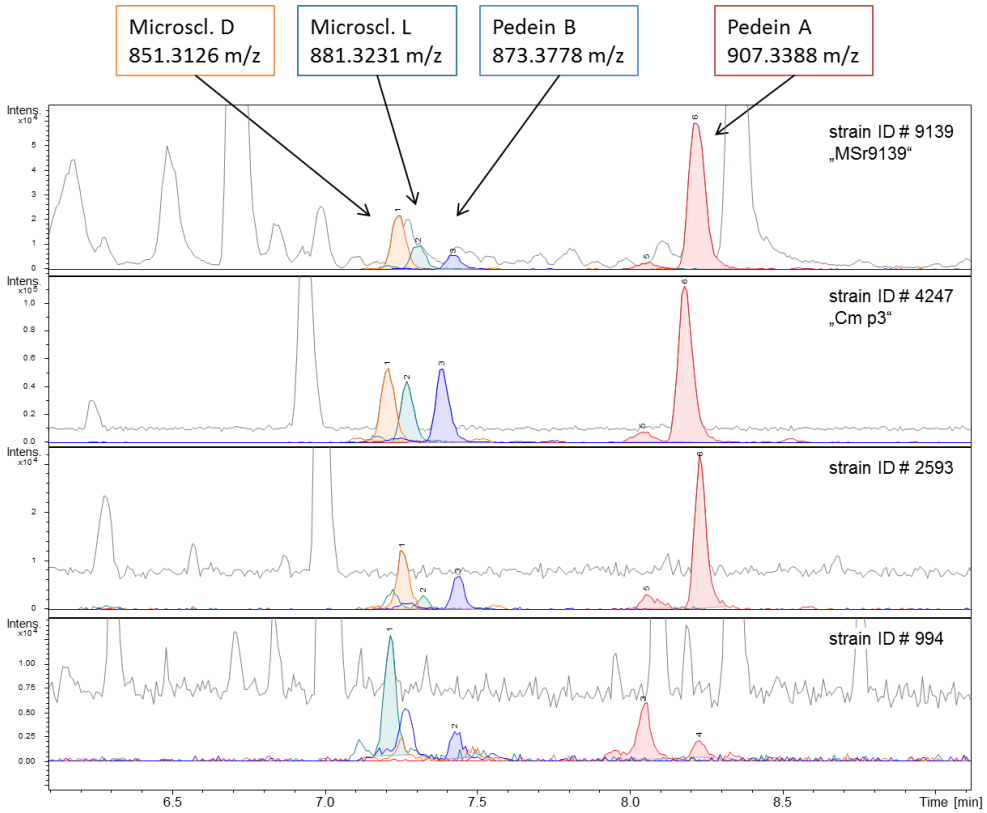


Figure S22: The known derivatives of microsclerodermin in four producer strains are highlighted by colored extracted ion chromatograms (EICs) overlaying the base peak chromatogram (BPC, grey). The producers are of the genus Jahnella and Chondromyces. Abundance of derivatives is related to the different producer strain. Identification of microsclerodermin derivatives is based on accurate m/z, isotope pattern fit, retention time and MS² pattern compared to the reference compounds.

3.6.6.1 LC-MS method for screening

All measurements were performed on a Dionex Ultimate 3000 RSLC system using a BEH C18, 100 x 2.1 mm, 1.7 µm dp column (Waters, Germany). Separation of 1 µl sample was achieved by a linear gradient from (A) H2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 600 µl/min and 45 °C. The gradient was 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 with initial conditions. 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 maXis 4G hr-ToF mass spectrometer (Bruker Daltonics, Germany) using the Apollo ESI source. Mass spectra were acquired in centroid mode ranging from 150 – 2500 m/z at 2 Hz scan rate.

3.6.7 Double Bond rearrangement

A retro-biosynthetic approach of the microsclerodermin side chain will initially result in a putative incorporation of a very uncommon C3 extender unit. This finding is owing to the double bond order which usually indicates the C2 extender units in between the double bonds. However, in microsclerodermin this does not hold true as the double bonds can rearrange in order to form a π-system conjugated to the phenyl ring. Taking this into account, the retro-biosynthetic proposal is based on a phenylacetate derived starter unit followed by C2 units, which is in agreement with conventional biosynthetic logic and the results from feeding experiments with labeled precursors. It remains elusive if this rearrangement is the result of sample extraction and workup or already takes place during/after biosynthesis *in vivo*.

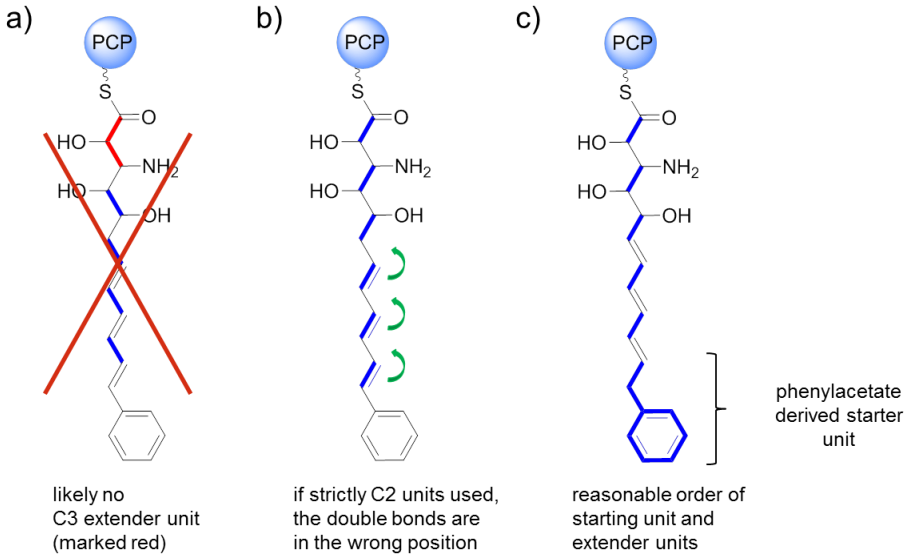


Figure S23: Retro-biosynthetic proposal for the microsclerodermin side chain. a) C3 extender unit is likely not incorporated during biosynthesis. b) A double bond rearrangement leads to C2 extender units. c) A reasonable biosynthetic proposal is starting with a phenylacetate starter followed by several C2 extender units.

3.6.8 Targeted inactivation of the *msc* locus in So ce38

Genetic modification of the So ce38 strain was accomplished according to a previously described protocol.[14] Putative single cross-over transformants were grown in H-Medium (0.2 % soybean meal, 0.2 % glucose, 0.8% starch, 0.2% yeast-extract, 0.1 % $CaCl_2 \cdot 2\ H_2O$, 0.1 % $MgSO_4 \cdot 7H_2O$, 8 mg/L FeEDTA, 50 mM HEPES, adjusted to pH 7.4 using 10N KOH) supplemented with hygromycin (100 µg/mL) and 1 % adsorber resin (XAD-16, Rohm & Haas) for at least 5 days at 180 rpm and 30 °C. Methanolic extracts of cells and the XAD resin were analyzed by LC-MS and compared with WT extracts. For isolation of chromosomal DNA the mutant strain was grown in M-medium containing 100 µg/mL hygromycin. The integration of the plasmid at the right position in the genome was verified by PCR as shown in supporting Figure S24, using mutant and wild-type DNA as templates.

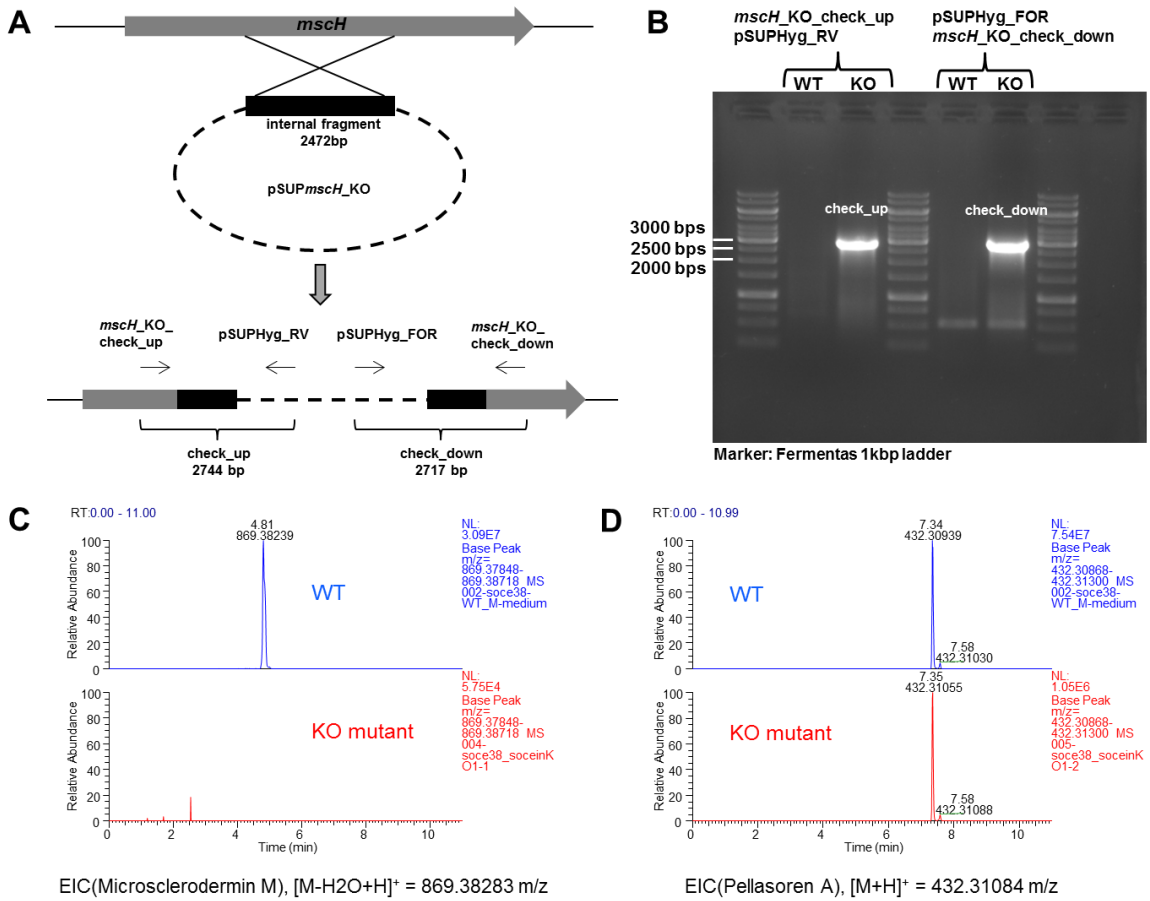


Figure S24: (A) Site of homologous recombination in So ce38. (B) Verification of KO-mutant So ce38::pSUPmscH_KO using PCR resulting in two amplicons "check_up" and "check_down". (C) Abolishment of microsclerodermin production as detected by comparative LC-MS analysis of wildtype and mutant strains. An extracted ion chromatogram for the most abundant [M-H_2O+H]$^+$ signal is shown. (D) Regular secondary metabolite production of the mutant strain was verified by identification of a known secondary metabolite in both wild-type and mutant strain.

Table S14: Primers used for amplification and verification of the microsclerodermin knock-out mutant in So ce38.

oligonucleotide	sequence
*mscH*_KO_for	GAT CCA GCG CTG GTT CCT CG
*mscH*_KO_for	ACT CGC CCT CGC GGA GGT TCT
pSUPHyg_fwd	ATG TAG CAC CTG AAG TCA GCC
pSUPHyg_rev	ACG CAT ATA GCG CTA GCA GC
*mscH*_KO_check_up	ACA ACT TCT TCG CGC TCG G
*mscH*_KO_check_down	TCG TCG TAC GAG AGC CGG

3.6.9 Feeding experiments using labeled precursors

3.6.9.1 Feeding in *S. cellulosum* So ce38

$^{15}N_2$-$^{13}C_4$-L-asparagine
(completely incorporated)
$\Delta m_{theor.}$ = 2 x ^{15}N and 4 x ^{13}C
= (2 x 0.997035 Da) + (4 x 1.003355 Da)
= 6.00749 Da
$\Delta m/z_{meas.}$ = 6.0075

^{15}N-$^{13}C_9$-L-phenylalanine:
(eight carbons incorparated)
8 x ^{13}C
$\Delta m_{theor.}$ = 8 x ^{13}C
= 8 x 1.003355 Da
= 8.02684 Da
$\Delta m/z_{meas.}$ = 8.0267

$^{13}C_6$-L-phenylalanine (ring label)
(six carbons incorporated)
6 x ^{13}C
$\Delta m_{theor.}$ = 6 x ^{13}C
= 6 x 1.003355 Da
= 6.02013 Da
$\Delta m/z_{meas.}$ = 6.02023

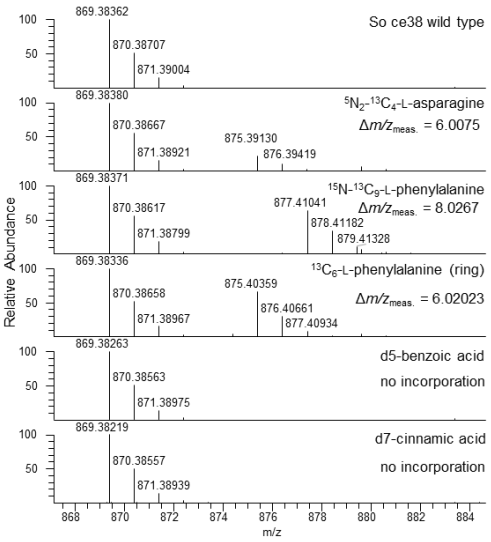


Figure S25: Isotopic peak pattern of the highly abundant [M-H2O+H]+ signal at 869.38283 m/z proves incorporation of labeled precursors into microsclerodermin M (So ce38). The observed mass shifts do perfectly fit to the heavy isotopes that were incorporated.

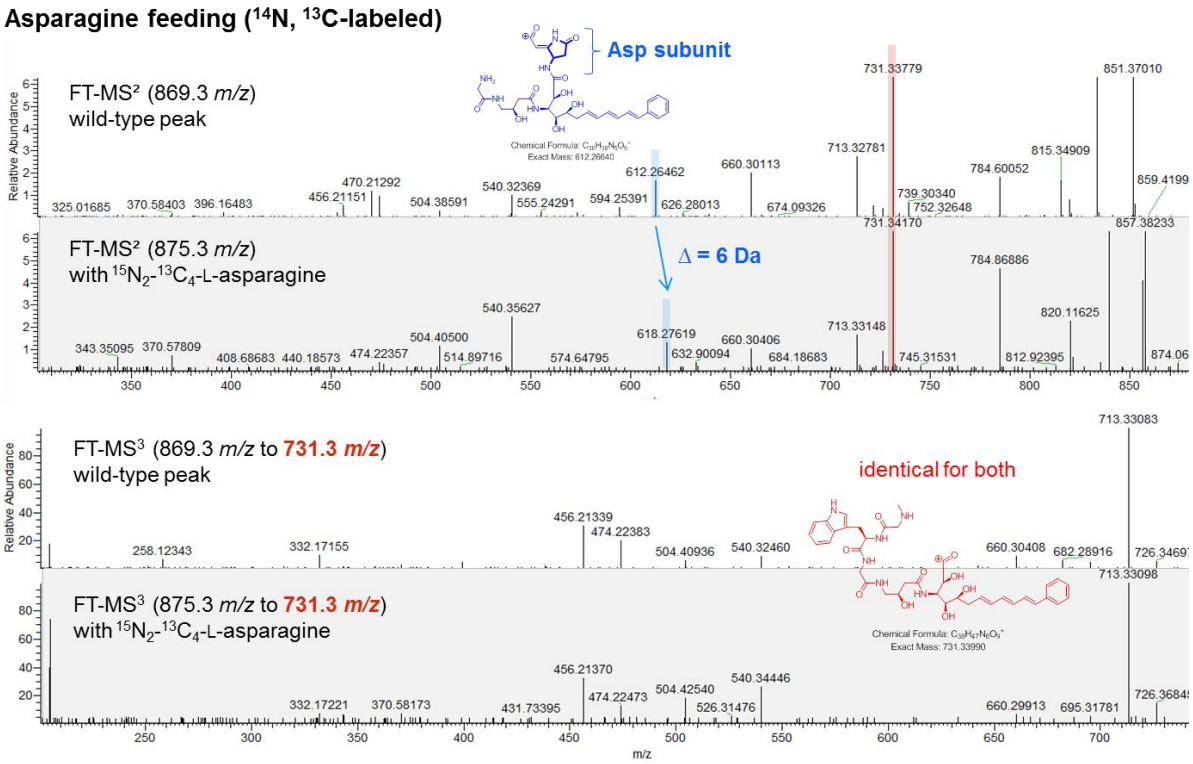


Figure S26: Incorporation of labeled L-asparagine subunit characterized by MS² and MS³ spectra acquisition. Data shows the feeding experiment in S.cellulosum So ce38. The most abundant fragment with 731.3 m/z is identical for the wild type peak and the enriched derivative (see MS³ data).

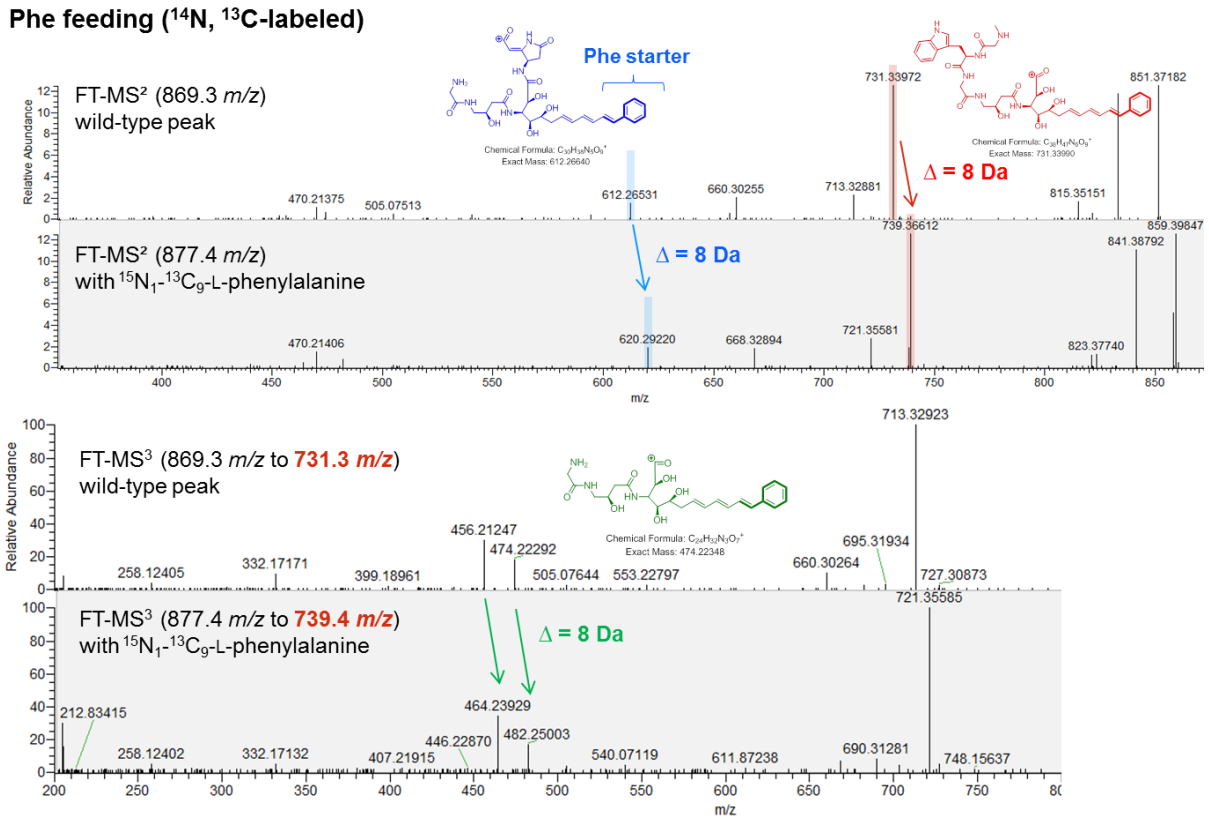


Figure S27: Incorporation of labeled phenylalanine subunit characterized by MS² and MS³ spectra acquisition. Data shows the feeding experiment in S.cellulosum So ce38. The mass difference of 8 Da is present in all the highlighted fragments indicating that phenylalanine is incorporated into the side chain.

Feeding in So ce38. Cultivation of So ce38 was performed in 25 mL H medium using 1 % (w/v) XAD-16. Labeled amino acids (5 mg) were dissolved in 250 mM HCl (320 µL) and sterile filtered. The labeled precursor stock (0.16 mg/mL) was added in three portions to the culture. 80 µL were added on first and third day followed by the remaining 160 µL on the fifth day of cultivation. Labeled benzoic acid and cinammic acid were dissolved in 80 µL DMSO and added in 2 x 20 µL and 40 µl portions to the respective culture. The cultures were harvested after 9 days and extracted with 2 x 25 mL methanol. The organic solvent was removed and the residuals dissolved in 500 µL methanol.

3.6.9.2 Feeding in *Jahnella* sp. MSr9139

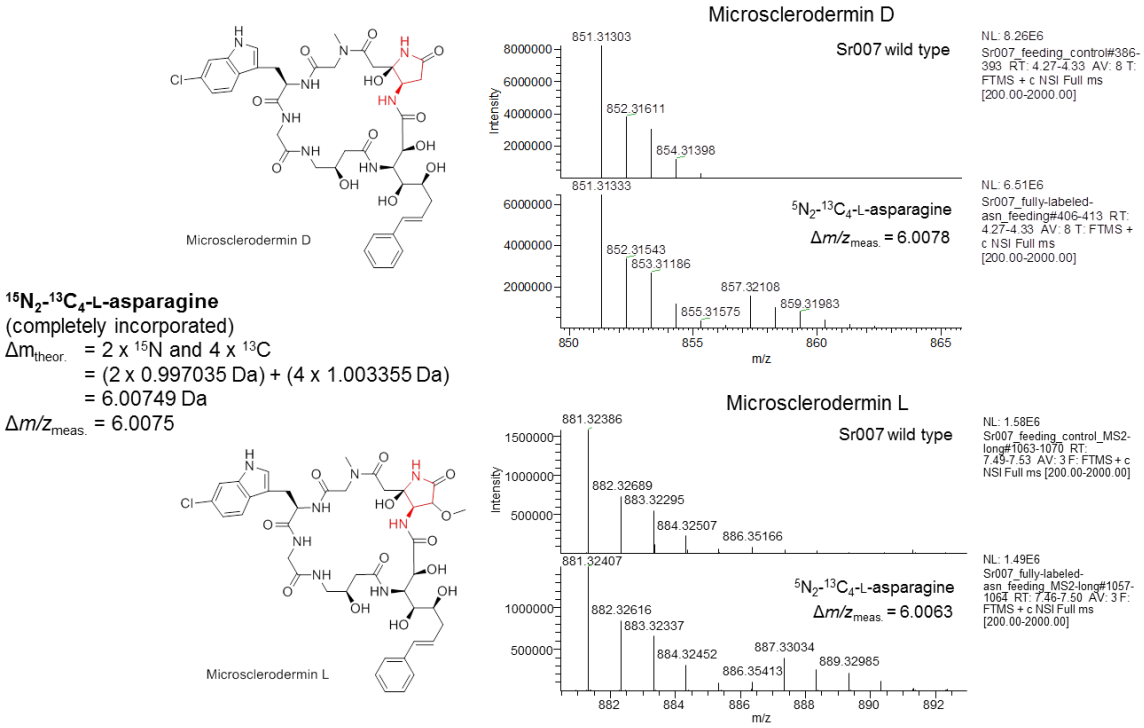


Figure S28: Isotopic peak pattern of the highly abundant $[M\text{-}H_2O+H]^+$ signals for microsclerodermin D (88 m/z) and L (881.32312 m/z) prove incorporation of labeled precursor in Jahnella sp. MSr9139. The observed mass shifts do perfectly fit to the heavy isotopes that were incorporated.

LC-MS analysis of feeding experiments. Measurements were performed on a Dionex Ultimate 3000 RSLC system using a Waters BEH C18, 50 x 2.1 mm, 1.7 µm d_p column by injection of two µl methanolic sample. Separation was achieved by a linear gradient with (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 600 µL/min and 45 °C. The gradient was initiated by a 0.3 min isocratic step at 5 % B, followed by an increase to 95 % B in 9 min to end up with a 1 min flush step at 95 % B before reequilibration with initial conditions. UV and MS detection were performed simultaneously. Coupling the HPLC to the MS was supported by an Advion Triversa Nanomate nano-ESI system attached to a Thermo Fisher Orbitrap. Mass spectra were acquired in centroid mode ranging from 200 – 2000 *m/z* at a resolution of R = 30000. The measurements were repeated to obtain high resolution CID data in order to support the experiment with annotation of the obtained fragments.

3.6.10 Bioactivity Testing

3.6.10.1 Bacterial Cultures

All microorganisms were handled under standard conditions recommended by the depositor. Overnight cultures of microorganisms were prepared in EBS medium (0.5 % peptone casein, 0.5 % proteose peptone, 0.1 % peptone meat, 0.1 % yeast extract; pH 7.0) or TSB medium (1.7 % peptone

casein, 0.3 % peptone soymeal, 0.25 % glucose, 0.5 % NaCl, 0.25 % K_2HPO_4; pH 7.3). The latter medium was used for *E. faecalis* and *S. pneumonia* cultures. Yeast and fungi were grown in Myc medium (1% phytone peptone, 1 % glucose, 50 mM HEPES, pH 7.0).

3.6.10.2 Microbial Susceptibility Assay (MIC)

Overnight cultures of microorganisms were diluted to OD_{600} 0.01 (bacteria) or 0.05 (yeast/fungi) in the respective growth medium. Serial dilutions of compounds were prepared as duplicates in sterile 96-well plates. The cell suspension was added and microorganisms were grown overnight on a microplate shaker (750 rpm, 30 °C or 37 °C). Growth inhibition was assessed by measuring the OD_{600} on a plate reader. MIC_{50} values were determined as average relative to respective control samples by sigmoidal curve fitting.

3.7 References

3.7.1 Main Text

(1) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2012**, *75*, 311–35.
(2) Wenzel, S. C.; Müller, R. *Mol. Biosyst.* **2009**, *5*, 567–74.
(3) Weissman, K. J.; Müller, R. *Bioorg. Med. Chem.* **2009**, *17*, 2121–36.
(4) Lane, A. L.; Moore, B. S. *Nat. Prod. Rep.* **2011**, *28*, 411–28.
(5) Winder, P. L.; Pomponi, S. A.; Wright, A. E. *Mar. Drugs* **2011**, *9*, 2643–82.
(6) Wink, J.; Müller, R. *Int. J. Med. Microbiol.* **2013**, in press.
(7) Hentschel, U.; Piel, J.; Degnan, S. M.; Taylor, M. W. *Nat. Rev. Microbiol.* **2012**, *10*, 641–54.
(8) Taylor, M. W.; Radax, R.; Steger, D.; Wagner, M. *Microbiol. Mol. Biol. Rev.* **2007**, *71*, 295–347.
(9) Bewley, C. A.; Faulkner, D. J. *Angew. Chem. Int. Ed. Engl.* **1998**, *37*, 2162–2178.
(10) Schmidt, E. W.; Obraztsova, A. Y.; Davidson, S. K.; Faulkner, D. J.; Haygood, M. G. *Mar. Biol.* **2000**, *136*, 969–977.
(11) Unson, M. D.; Faulkner, D. J. *Experientia* **1993**, *49*, 349–353.
(12) Elshahawi, S. I.; Trindade-Silva, A. E.; Hanora, A.; Han, A. W.; Flores, M. S.; Vizzoni, V.; Schrago, C. G.; Soares, C. A.; Concepcion, G. P.; Distel, D. L.; Schmidt, E. W.; Haygood, M. G. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, E295–304.
(13) Lin, Z.; Torres, J. P.; Ammon, M. A.; Marett, L.; Teichert, R. W.; Reilly, C. A.; Kwan, J. C.; Hughen, R. W.; Flores, M.; Tianero, M. D.; Peraud, O.; Cox, J. E.; Light, A. R.; Villaraza, A. J. L.; Haygood, M. G.; Concepcion, G. P.; Olivera, B. M.; Schmidt, E. W. *Chem. Biol.* **2013**, *20*, 73–81.
(14) Kwan, J. C.; Donia, M. S.; Han, A. W.; Hirose, E.; Haygood, M. G.; Schmidt, E. W. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 20655–60.
(15) Wenzel, S. C.; Müller, R. In *Comprehensive Natural Products II, Vol. 2: Natural Products Structural Diversity-II Secondary Metabolites: Sources, Structures and Chemical Biology*; Moore, B. S.; Crews, P., Eds.; Elsevier, 2010; pp 189-222.
(16) Crews, P.; Manes, L. V; Boehler, M. *Tetrahedron Lett.* **1986**, *27*, 2797–2800.
(17) Zabriskie, T. M.; Klocke, J. A.; Ireland, C. M.; Marcus, A. H.; Molinski, T. F.; Faulkner, D. J.; Xu, C.; Clardy, J. *J. Am. Chem. Soc.* **1986**, *108*, 3123–3124.
(18) Kunze, B.; Jansen, R.; Sasse, F.; Höfle, G.; Reichenbach, H. *J. Antibiot.* **1995**, *48*, 1262–6.
(19) Irschik, H.; Trowitzsch-Kienast, W.; Gerth, K.; Höfle, G.; Reichenbach, H. *J. Antibiot.* **1988**, *41*, 993–8.
(20) Frincke, J. M.; Faulkner, D. J. *J. Am. Chem. Soc.* **1982**, *104*, 265–269.
(21) Erickson, K. L.; Beutler, J. A.; Cardellina, J. H.; Boyd, M. R. *J. Org. Chem.* **1997**, *62*, 8188–8192.
(22) Kunze, B.; Jansen, R.; Sasse, F.; Höfle, G.; Reichenbach, H. *J. Antibiot.* **1998**, *51*, 1075–80.
(23) Hoffmann, H.; Haag-Richter, S.; Kurz, M.; Tietgen, H. Bengamide derivatives, method for the production thereof and use thereof for the treatment of cancer. WO2005044803 A1, 2005.
(24) Johnson, T. A.; Sohn, J.; Vaske, Y. M.; White, K. N.; Cohen, T. L.; Vervoort, H. C.; Tenney, K.; Valeriote, F. A.; Bjeldanes, L. F.; Crews, P. *Bioorg. Med. Chem.* **2012**, *20*, 4348–55.
(25) Kunze, B.; Böhlendorf, B.; Reichenbach, H.; Höfle, G. *J. Antibiot.* **2008**, *61*, 18–26.
(26) Bewley, C. A.; Debitus, C.; Faulkner, D. J. *J. Am. Chem. Soc.* **1994**, *116*, 7631–7636.
(27) Schmidt, E. W.; John Faulkner, D. *Tetrahedron* **1998**, *54*, 3043–3056.
(28) Qureshi, A.; Colin, P. L.; Faulkner, D. J. *Tetrahedron* **2000**, *56*, 3679–3685.
(29) Zhang, X.; Jacob, M. R.; Ranga Rao, R.; Wang, Y. H.; Agarwal, A. K.; Newman, D. J.; Khan, I. A.; Clark, A. M.; Li, X. C. *Res. Rep. Med. Chem.* **2012**, *2012*, 7–14.
(30) Zhang, X.; Jacob, M. R.; Ranga Rao, R.; Wang, Y. H.; Agarwal, A. K.; Newman, D. J.; Khan, I. A.; Clark, A. M.; Li, X. C. *Res. Rep. Med. Chem.* **2013**, 9.
(31) Simister, R. L.; Deines, P.; Botté, E. S.; Webster, N. S.; Taylor, M. W. *Environ. Microbiol.* **2012**, *14*, 517–24.
(32) Garcia, R. O.; Krug, D.; Müller, R. *Meth. Enzymol.* **2009**, *458*, 59–91.

(33) Sambrook, J.; Russell, D. W. *Molecular cloning: A laboratory manual*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2001.
(34) Kopp, M.; Irschik, H.; Gross, F.; Perlova, O.; Sandmann, A.; Gerth, K.; Müller, R. *J. Biotechnol.* **2004**, *107*, 29–40.
(35) Garcia, R.; Gerth, K.; Stadler, M.; Dogma, I. J.; Müller, R. *Mol. Phylogenet. Evol.* **2010**, *57*, 878–87.
(36) Garcia, R.; Müller, R. In *The Prokaryotes: Deltaproteobacteria and Epsilonproteobacteria*; Rosenberg, E.; DeLong, E. F.; Lory, S.; Stackebrandt, E.; Thompson, F., Eds.; Springer: Heidelberg, 2014, in press.
(37) Jahns, C.; Hoffmann, T.; Müller, S.; Gerth, K.; Washausen, P.; Höfle, G.; Reichenbach, H.; Kalesse, M.; Müller, R. *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 5239–43.
(38) Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; Breitling, R. *Nucleic Acids Res.* **2011**, *39*, W339–46.
(39) Butcher, R. A.; Schroeder, F. C.; Fischbach, M. A.; Straight, P. D.; Kolter, R.; Walsh, C. T.; Clardy, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1506–9.
(40) Kusebauch, B.; Busch, B.; Scherlach, K.; Roth, M.; Hertweck, C. *Angew. Chem. Int. Ed. Engl.* **2010**, *49*, 1460–4.
(41) Erol, O.; Schäberle, T. F.; Schmitz, A.; Rachid, S.; Gurgui, C.; El Omari, M.; Lohr, F.; Kehraus, S.; Piel, J.; Müller, R.; König, G. M. *ChemBioChem* **2010**, *11*, 1253–65.
(42) Taft, F.; Brünjes, M.; Knobloch, T.; Floss, H. G.; Kirschning, A. *J. Am. Chem. Soc.* **2009**, *131*, 3812–3.
(43) Magarvey, N. A.; Beck, Z. Q.; Golakoti, T.; Ding, Y.; Huber, U.; Hemscheidt, T. K.; Abelson, D.; Moore, R. E.; Sherman, D. H. *ACS Chem. Biol.* **2006**, *1*, 766–79.
(44) Tillett, D.; Dittmann, E.; Erhard, M.; von Döhren, H.; Börner, T.; Neilan, B. A. *Chem. Biol.* **2000**, *7*, 753–64.
(45) Röttig, M.; Medema, M. H.; Blin, K.; Weber, T.; Rausch, C.; Kohlbacher, O. *Nucleic Acids Res.* **2011**, *39*, W362–7.
(46) Plaza, A.; Bifulco, G.; Masullo, M.; Lloyd, J. R.; Keffer, J. L.; Colin, P. L.; Hooper, J. N. A.; Bell, L. J.; Bewley, C. A. *J. Org. Chem.* **2010**, *75*, 4344–55.
(47) Mujika, J. I.; Lopez, X.; Mulholland, A. J. *Org. Biomol. Chem.* **2012**, *10*, 1207–18.
(48) Gaitatzis, N.; Silakowski, B.; Kunze, B.; Nordsiek, G.; Blöcker, H.; Höfle, G.; Müller, R. *J. Biol. Chem.* **2002**, *277*, 13082–90.
(49) Yadav, G.; Gokhale, R. S.; Mohanty, D. *PLoS Comput. Biol.* **2009**, *5*, e1000351.
(50) Kozbial, P. Z.; Mushegian, A. R. *BMC Struct. Biol.* **2005**, *5*, 19.

3.7.2 Supporting Information

(1) Delcher, A. L.; Bratke, K. A.; Powers, E. C.; Salzberg, S. L. *Bioinformatics* **2007**, *23*, 673–9.
(2) Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; Breitling, R. *Nucleic Acids Res.* **2011**, *39*, W339–46.
(3) Yadav, G.; Gokhale, R. S.; Mohanty, D. *Journal of Molecular Biology* **2003**, *328*, 335–363.
(4) Ziemert, N.; Podell, S.; Penn, K.; Badger, J. H.; Allen, E.; Jensen, P. R. *PLoS ONE* **2012**, *7*, e34064.
(5) Caffrey, P. *ChemBioChem* **2003**, *4*, 654–7.
(6) Kwan, D. H.; Tosin, M.; Schläger, N.; Schulz, F.; Leadlay, P. F. *Org. Biomol. Chem.* **2011**, *9*, 2053–6.
(7) Röttig, M.; Medema, M. H.; Blin, K.; Weber, T.; Rausch, C.; Kohlbacher, O. *Nucleic Acids Res.* **2011**, *39*, W362–7.
(8) Kozbial, P. Z.; Mushegian, A. R. *BMC Struct. Biol.* **2005**, *5*, 19.
(9) Schmidt, E. W.; John Faulkner, D. *Tetrahedron* **1998**, *54*, 3043–3056.
(10) Qureshi, A.; Colin, P. L.; Faulkner, D. J. *Tetrahedron* **2000**, *56*, 3679–3685.
(11) Kunze, B.; Böhlendorf, B.; Reichenbach, H.; Höfle, G. *J. Antibiot.* **2008**, *61*, 18–26.
(12) Bewley, C. A.; Debitus, C.; Faulkner, D. J. *J. Am. Chem. Soc.* **1994**, *116*, 7631–7636.
(13) Fujii, K.; Ikai, Y.; Oka, H.; Suzuki, M.; Harada, K. *Analytical Chemistry* **1997**, *69*, 5146–5151.
(14) Kopp, M.; Irschik, H.; Gross, F.; Perlova, O.; Sandmann, A.; Gerth, K.; Müller, R. *J. Biotechnol.* **2004**, *107, 29-40.*

Chapter 4 – Secondary Metabolomics

Myxobacterial Secondary Metabolomics: Contributions to a Comprehensive Screening Workflow

4 Secondary Metabolomics

4.1 Introduction

Natural products from microorganisms are a valuable source for pharmaceutical lead structures.[1] Novel bioactive scaffolds are urgently needed for the development of new drugs to overcome the increasing antibiotic resistance amongst pathogenic bacteria.[2] A promising approach for the discovery of new chemical entities is to search for natural products in formerly underexploited sources, following the notion that chemical diversity comes along with phylogenetic diversity. In the long run, the main challenge is finding new natural products and isolating substantial amounts of these for full characterization and bioactivity testing. However, detection of potentially new compounds and their classification as novel scaffolds comes first. In addition, natural product discovery workflows are increasingly linked to genome mining approaches to support the identification of new secondary metabolites (Figure 1); a trend which certainly requires reliable analytical data to draw meaningful conclusions.

Traditional activity-guided isolation efforts are efficient in finding new bioactive molecules albeit this methodology suffers from being limited to a small set of assays and indicator strains.[3] Especially complex assays or those based on high-risk pathogenic bacteria of the ESKAPE group (see Chapter 1) are rarely used for routine screening. As a consequence, compounds with potent activities are frequently found after isolation rather than being the driving force behind the isolation. In summary, all efforts

point toward finding compounds that are new *and* active whereas it does not matter what was revealed first, activity or structural novelty.

The crucial part of compound identification in natural product research is usually based on analytical techniques like LC-MS as these are able to cope with the complexity of biological samples. Utilizing LC-MS or LC-MS/MS analysis provides a powerful and well established approach to identify known compounds based on reference substance databases. This process of dereplication facilitates the search for new secondary metabolites as the risk of re-isolating a known compound is thereby reduced.[4] However, targeted LC-MS workflows are limited to known compounds whereas methods addressing the problem of specifically revealing new compounds are still on their advent. Several computational techniques have recently emerged that use MS/MS data for classification of compounds[5], such as fragmentation trees[6,7], cyclic peptide de novo sequencing[8,9], and spectral network construction[10,11]. In addition, statistical data mining approaches were used to characterize knock-out mutants by principal component analysis (PCA)[12] or they served as a method to prioritize strains based on crude extract measurements in order to identify those strains that are the least similar to others within a given sample set.[13,14]

It is quite obvious that results obtained from these approaches critically depend on the coverage and quality of LC-MS(/MS) data available for a given sample set. This basic presumption applies particularly to secondary metabolites that are easily missed in a complex matrix owing to their low abundance or because of the overwhelming number of accompanying signals at the same retention time. Indeed, LC-MS profiles from microbial producers are frequently overloaded with large numbers of intense signals derived from highly abundant matrix components present in a sample. This can become a problem when fragmentation data is needed, e.g. in an untargeted approach where auto-MS² settings usually follow an intensity-based precursor selection. Unsupervised acquisition of MS/MS data will end up with high numbers of matrix components being fragmented while supposedly important peaks may be missed due to their lower abundance compared to matrix peaks. Hence, in this study we aim to improve LC-MS/MS analysis towards a reasonable MS/MS precursor selection in order to focus on compounds that are actually produced *de novo* by a bacterial strain, as within this selection of compounds chances are best to discover new and bioactive secondary metabolites.

This work demonstrates how statistical data mining in combination with subsequent scheduled MS/MS acquisition can help to highlight and identify mass spectral features that are related to bacterial metabolism. In particular, these features are defined by retention time, *m/z* value and a detection response measure (i.e. peak intensity or peak area). The whole method is based on the presumption that these features originate from distinct compounds which are detected by the mass spectrometer in the course of chromatographic separation. Hence, such an analysis results in a list of putative microbial metabolites (represented by molecular features) that helps to address two important issues, namely the

identification of potentially interesting peaks in loaded regions of the chromatogram plus the acquisition of the corresponding MS/MS spectra irrespective of matrix compounds that cover the compounds of interest.

We exemplify the devised workflow for myxobacteria which usually produce complex crude extracts, a consequence of the complex media needed to support growth of many myxobacterial species and the minimalistic workup procedures used for extract preparation. The methodology described here in combination with in-house database queries and directed bioactivity screening of fractionated crude extracts enabled us to generate a well-defined set of candidate signals with their MS/MS spectra that is focused on relevant myxobacterial compounds.

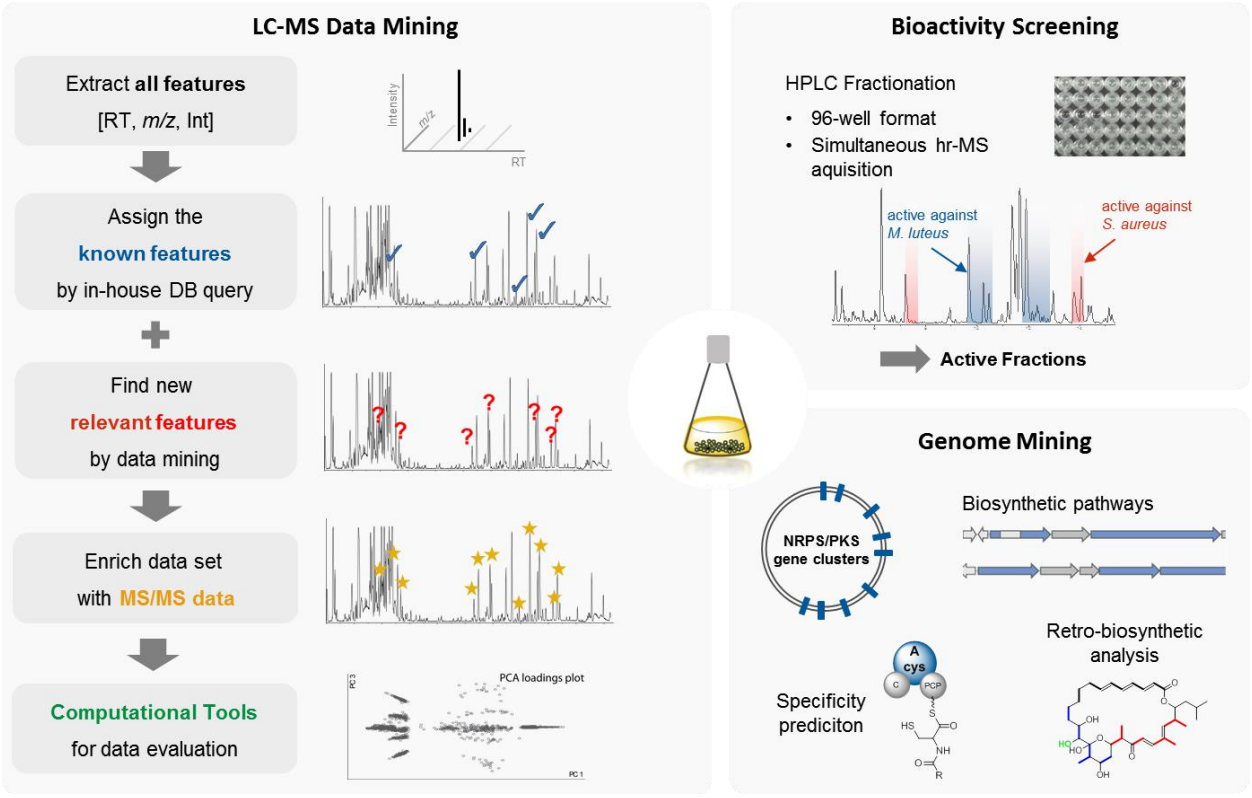


Figure 1: Schematic overview of three important approaches supporting and driving natural product research: LC-MS analysis, bioactivity screening, and genome mining. The combination of these techniques helps to uncover new natural products from crude bacterial extracts.

4.2 LC-MS Performance Evaluation and System Suitability

The suitability of an analytical platform for a given task critically depends on the system's performance. In terms of dereplication, which is regarded as a method of targeted metabolomics, retention time stability, an accurate *m/z* value with less than 5 ppm error, and a flawless detection of isotope intensities is essential to facilitate successful identification of known compounds. Reliability is even more crucial for approaches of untargeted secondary metabolomics as no references are available that indicate systemic errors. Thus, any analytical platform has to fulfill pre-defined performance criteria

tailored to the specific requirements. This performance needs to be approved on a regular basis to confirm that the system can deliver the expected quality and that measurements must not be conducted if the system fails to meet specified criteria. In this regard, an analytical platform was established for the analyses of myxobacterial extracts aiming to serve as the basis for a comprehensive screening program for both, targeted methods of dereplication and for untargeted secondary metabolomics approaches.

The system is based on chromatographic separation employing reversed phase chromatography with C18-modified particles of the sub-2 µm class combined with mass spectrometric data acquisition using an QqTOF machine in positive ESI mode. An unspecific linear gradient is used to separate secondary metabolites as they are found at any elution strength (Figure 2). The compounds of interest are mostly polyketides, peptidic compounds, or hybrids thereof that are usually further modified by various tailoring enzymatic steps such as oxidation, methylation, chlorination and others (see Chapter 1). Although some of the identified compounds feature ionic moieties such as carboxylic acids, amines, phosphate groups, or sulfate moieties, retention is always high enough to separate these compounds with reversed phase chromatography.

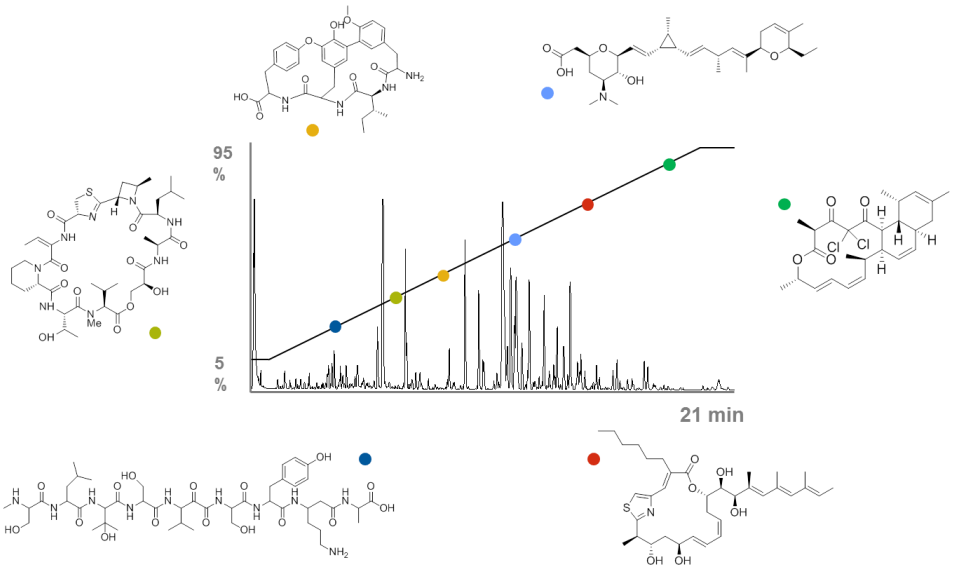


Figure 2: Myxobacterial secondary metabolites are found at various retention times when they are separated with an unspecific gradient using reversed phase chromatography. A large scale screening project must not be specific as too many compounds would be missed.

A certain bias is always introduced by the methods of extraction and measurement, but this is to some degree volitional as the workflow thereby focusses on the detection of secondary metabolites. In particular, secondary metabolites are enriched on polymeric adsorber resin during bacterial growth and eluted afterwards by methanol or acetone. Compounds of primary metabolism are poorly adsorbed when using this approach. At the same time, the LC-MS method is based on RP-HPLC using a C18-modified stationary phase thereby addressing secondary metabolites rather than small ionic

metabolites of primary metabolism. In summary, the analytical platform is suitable for almost all known myxobacterial secondary metabolites. In addition free fatty acids, lipids and steroids are frequently detected by this method although it is conceivable that other types of analysis may be better for certain classes of metabolites and some metabolites may even circumvent detection when this method is used.

All known myxobacterial compounds were measured on this analytical platform resulting in a large reference set of retention time, observed adducts, and MS^2 spectra deposited in an in-house database. The "Myxobase" is a proprietary database that aims to be comprehensive with respect to myxobacterial research by aggregating all relevant information about strains and compounds together with previously acquired analytical information and bioactivity testing results. At the time of this work the Myxobase comprised analytical information of more than 900 compounds and approximately 2500 LC-MS data sets derived from myxobacterial extracts. Pure compounds and extracts were measured using standardized conditions and subsequently served as input for two different processing and analysis workflows:

1) *Targeted screening*
   Known compounds are identified and annotated in LC-MS data of extracts by matching against compound information that is deposited in the Myxobase.

2) *Molecular feature "warehousing"*
   Each LC-MS run is subjected to comprehensive molecular feature extraction and the resulting set of molecular features is deposited in Myxobase.

These feature sets, generated in a largely unbiased way, can represent the informational content as found in the respective raw data files. A putatively new compound can be searched within this huge data set, thereby offering the chance to find other strains that produce the same compound in higher yield. It goes without saying that a large-scale analytical framework like this requires a well-monitored and validated analytical system to rely on. Hence, several important cornerstones of such a screening platform in natural product research will be addressed in the following.

4.2.1 LC-MS Robustness

Retention time is an important analytical readout as it is orthogonal to mass spectral information and thus mandatory to use when conducting LC-MS based screening of microbial extracts. With respect to this, reproducible retention times are essential to support identification of known compounds when utilizing database queries (Chapter 4.3). Similarly, the data mining approaches described later in Chapter 4.5 of this work

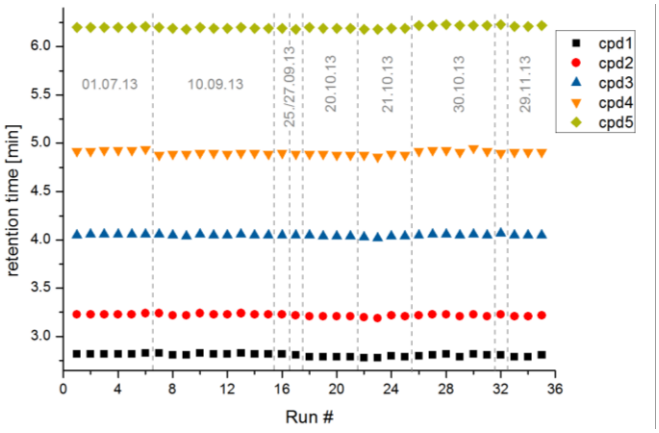


Figure 3: Long term retention time stability. Plot of test mix retention times spanning five months and 2800 samples measured during this time. Shown are only the test mix samples that were part of this project's sequence lists.

critically depends on reproducible retention times. For this reason, daily test mix runs were employed to ensure retention time stability. Excellent retention time stability was maintained throughout this project as depicted in Figure 3. In line with the long-term stability, Table 1 shows the very good retention time stability for a series of measurements (120 LC-MS runs) spanning 48 hours of data acquisition which is the basis for the data mining workflow of Chapter 4.5,

Table 1: Test mix performance data (N = 9, main data set from 2013-09-10).

Peak	*m/z*	RT [min]	RT RSD	Area RSD	Intensity RSD	*m/z* error [ppm]
1	279.091023	2.82	0.3%	6.0%	4.6%	-1.84 ± 0.49
2	285.020751	3.23	0.2%	10.4%	6.9%	-3.26 ± 0.46
3	311.080852	4.05	0.2%	8.1%	6.7%	-1.61 ± 0.53
4	278.190326	4.89	0.1%	2.8%	3.0%	-0.75 ± 0.51
5	314.138685	6.19	0.1%	4.6%	6.9%	-2.48 ± 0.49

In addition, a myxobacterial crude extract was measured after each test mix. Being a more complex quality control sample, this extract covers a broad range of known secondary metabolites and thereby provides an overview on the system's *m/z* accuracy, mSigma values and area value deviations. The "mSigma" value is a measure for the quality of the isotope intensity distribution. It compares the experimental isotope intensity with the calculated isotope intensity of a known compound (based on the known sum formula). Acceptable values cover values below 30 mSigma. The results presented in Table 2 highlight the robustness of the analytical platform with most part of the *m/z* errors below 3 ppm and very good mSigma values. Chromatographic peak area deviations of around 10 % RSD are in line with typical values when using an ESI ion source.

Table 2: System repeatability study using known compounds in a crude extract of So ceGT47. Samples were measured in the course of a 48 h sequence just after the test mix samples of Figure 3 (N=9).

Compound name	Myxobase ID	RT [min]		*m/z* error [ppm]		mSigma		Area	Int.
		mean	STDEV	mean	STDEV	mean	STDEV	RSD	RSD
Cyclic Tyr-Pro	2425	2.41	0.01	0.90	1.72	6.5	3.3	9.3%	7.6%
SH-Myxo-373_1	2824	4.29	0.00	0.70	1.16	13.3	2.1	11.5%	11.7%
SH-Cm7-407	2889	4.73	0.01	0.59	1.37	12.9	4.8	11.3%	11.7%
Lipothiazole A	2572	9.17	0.01	0.61	1.91	15.8	1.2	6.8%	5.1%
Lipothiazole B	2702	9.73	0.01	0.38	0.72	20.9	2.8	9.1%	8.4%
Lipothiazole D	2704	9.91	0.01	-1.10	1.89	13.8	1.4	11.7%	10.9%
Lipothiazole C	2703	9.94	0.01	-1.24	2.12	11.6	1.6	10.7%	10.7%
Ambruticin VS-5	224	10.69	0.02	-0.13	1.69	9.6	9.6	9.8%	12.3%
Ambruticin VS-3	83	10.96	0.02	0.08	0.92	18.1	3.3	11.2%	9.5%
Ambruticin VS-1	221	11.17	0.02	-0.04	0.68	20.8	2.6	11.6%	9.0%
Thuggacin B	305	12.40	0.01	0.88	0.88	23.6	3.0	12.4%	9.9%
Ambruticin S	216	12.84	0.01	-0.94	4.01	7.6	3.4	8.2%	6.6%
Ambruticin F	311	13.20	0.01	0.51	1.48	12.0	3.7	3.5%	2.7%
Thuggacin A	75	13.46	0.01	0.10	0.73	22.2	3.2	5.0%	5.6%
Thuggacin C	528	14.06	0.01	-0.21	1.49	17.7	1.4	21.1%	19.5%

These results well justify the conclusion that the system is highly stable and able to deliver reproducible chromatographic separations alongside with accurate and reliable mass spectrometric data acquisition.

4.2.2 Biological Reproducibility

Sample to sample reproducibility with respect to cultivation and extract preparation is an important aspect when trying to identify differences between sample groups. Differential analysis requires as a prerequisite that significant variations are largely absent within a group but are detectable in between groups. Using seven different cultivation conditions – all produced in biological triplicates – we can evaluate whether cultivation and preparation of crude extracts is sufficiently reproducible. An overlay chromatogram of three independently prepared extracts reveals the biological reproducibility (Figure 4). The overlay display is split into four retention time ranges to allow better comparison of the single chromatogram traces.

For most part of a LC-MS run, the peaks of a base peak chromatogram (BPC) are congruent for biological replicates of an extract. However, at very late retention times with high organic content the BPC is frequently different (Figure 4 D, blue trace). The compounds causing the BPC in this region remain unknown although there is certain likelihood that lipids play a role. The important conclusion of these triplicate measurements is that the number of outliers was remarkably low within the extracts used in this study. Only one extract showed additional peaks in the chromatogram while for two extracts some peak areas varied gradually. Altogether, this low biological variance is quite remarkable as strains were grown and extracted independently. A single crude extract may be different for unknown reasons, e.g.

one of the six So ceGT47 in P medium shows a few additional huge peaks. The reason for this may be a contamination or irregular growth conditions but fortunately, deviations comprising only a few peaks in single samples do not affect the statistics-based workflow.

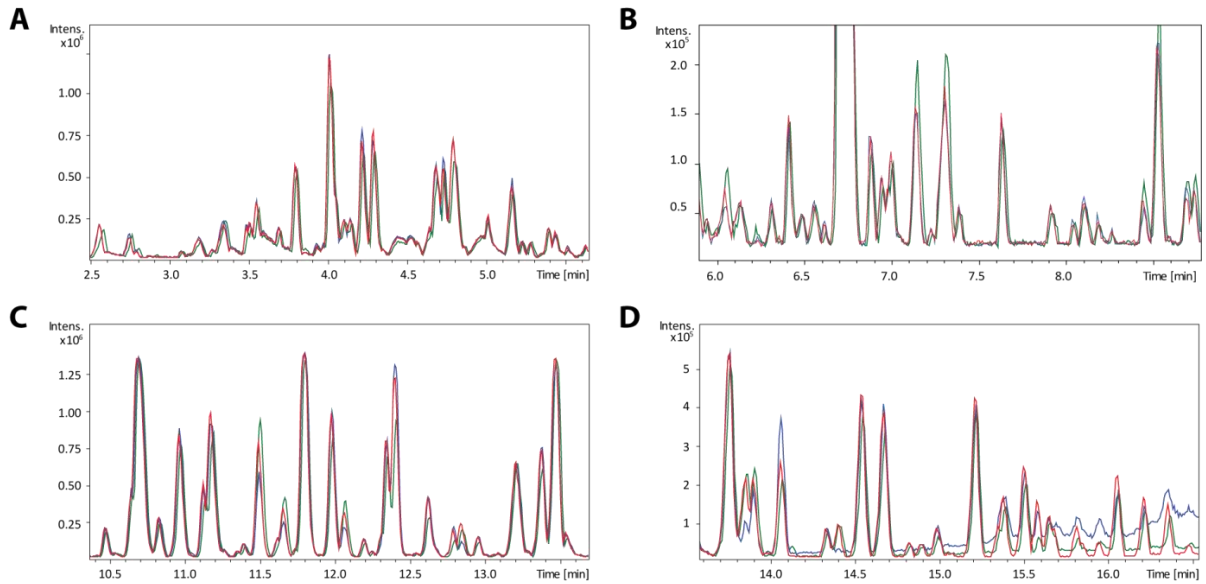


Figure 4: Overlay display of the base peak chromatograms of three So ceGT47 crude extracts derived from P-medium cultures (red, blue, green). The different panels feature several retention time windows, which were zoomed for a better comparison.

4.2.3 Matrix effects in crude extracts

The effect of a complex sample matrix on the detection of known compounds was evaluated. This analysis is of interest as estimations of production rates are often based on measuring the peak area of a compound in crude extracts. This bears a risk as the influence of ion suppression is not predictable. In particular, ion suppression for a compound is caused by other compounds eluting at the same retention time. Different complex cultivation media will certainly cause different peaks within a chromatographic separation. As these peaks can elute at different retention times, it becomes clear that suppression effects are related to retention time and the medium used. Hence, we set out to quantify the effect of ion suppression for three known myxobacterial compounds eluting at the beginning, the middle and the end of the standard chromatographic separation (Figure 5 B). These compounds are not naturally present in extracts of So ceGT47 but were spiked at different concentrations into the crude extracts derived from cultivation of this strain using A, P, H, or M medium, respectively. In addition to spiking crude extracts, a dilution series in pure methanol was used as reference. Area values derived from spiked crude extracts were set in relation to the reference samples to receive a measure for the matrix-induced suppression. A value of 100 % suppression means total loss of peak detection (Figure 5). While concentrations of 0.1 µg/µL are readily identified in reference samples, the same concentration is not detectable in crude extracts. This suppression effect is very strong for the highly lipophilic chlorotonil,

whereas sorangicin is only slightly affected. Chlorotonil elutes at the very end of the chromatogram and is heavily suppressed although almost no other peaks are visible. There may be apolar components not detected by ESI-MS which strongly suppress ionization in this chromatographic region. An influence of the cultivation medium is obvious when M-medium and A-medium are compared with respect to the detection of chlorotonil (Figure 5 D).

Overall, numerous influences add up to a suppression that cannot be foreseen. As a consequence, yield estimations in crude extracts relying on LC-MS measurements should be regarded as rough estimates as long as they are based on a simple calibration curve that was created with pure samples, e.g. methanolic dilutions of a compound. Concentration values obtained using such simple calibrations are most likely an underestimate of the real value. Thus, exact ways to quantify the production in crude extracts by LC-MS are limited to standard addition methods or methods considering compound-specific suppression factors.

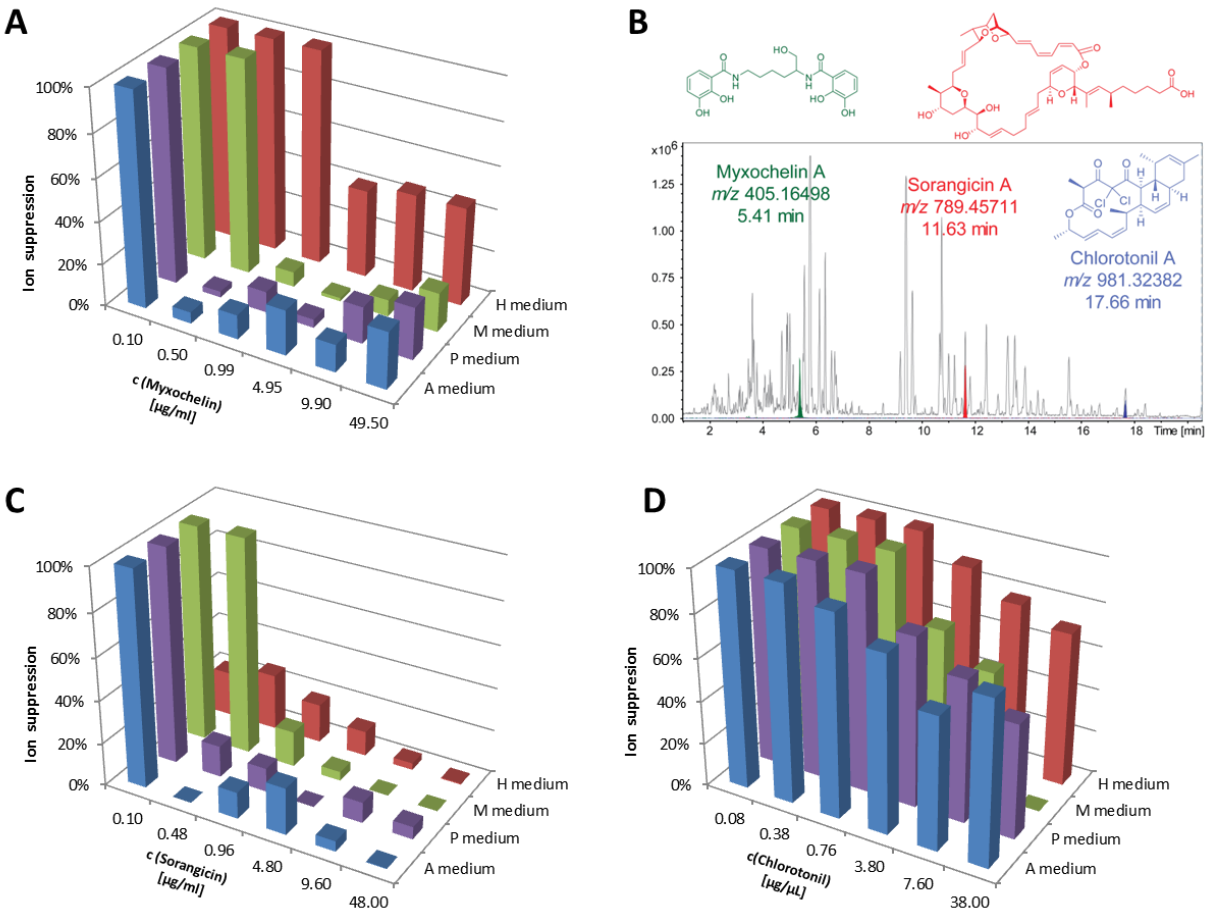


Figure 5: ***A, C, D*** *Suppression of signal intensities for selected spiked myxobacterial compounds. Values refer to the change in intensity compared to a methanolic sample of equal concentration, i.e. 100 % is total loss of detection. Concentration values are corrected for compound purity as determined independently.* ***B*** *So ceGT47 M medium crude extract spiked with 3-compound mix to a final concentration of 50 µg/mL.*

This finding is also of importance when different media are compared as an increase or decrease of a peak area is not necessarily owing to a change in production rate. The suppression effect should be considered when statistical mining approaches aim toward the analysis of a sample set that was derived from extracts grown with different media. However, all comparative analyses presented in this work are restricted to samples grown in the same medium; hence, no problems should arise by the suppression effect in this study.

4.3 Identification of Known Compounds

Research at the Helmholtz Center for Infection Research, Braunschweig, Germany and at the Helmholtz Institute for Pharmaceutical Research Saarland, Saarbrücken, Germany yielded more than 900 myxobacterial compounds of around 140 compound classes, with a notable unpublished number (as of 01/2014). With few exceptions these compounds are listed in our proprietary database "Myxobase" with analytical information such as retention time, typical adducts and neutral loss fragments that are usually observed (Chapter 4.2). By having this information available, we can query the database and identify already known compounds in a myxobacterial extract. The database query is performed with the Bruker TargetAnalysis™ software based on a reference list from the Myxobase containing the necessary analytical information. A LC-MS run is checked for presence of the compounds as listed in this reference. Scoring of hits is based on retention time accuracy, *m/z* accuracy and the isotope intensity distribution of each isotope pattern (mSigma value). The isotope pattern is a function of the sum formula which is provided with the reference list and comparing the theoretical isotope pattern with the measured one gives an orthogonal analytical readout, thereby adding another level of confidence. This readout is especially important for compounds exceeding molecular weights of 500 g/mol like it is the case for many natural products.[15] In addition, commonly observed adducts or neutral loss peaks of compounds are also deposited in the database. Their identification further increases the true positive hit rate. Applying this workflow to a myxobacterial extract enables detection of known compounds which results in a fully annotated chromatogram for a convenient manual evaluation if necessary (Figure 6).

Knowledge of which compounds are present in a bacterial extract allows prediction of the extract's bioactivity profile. Especially activities that do not match the known compounds are of interest as they point toward new chemical entities. While this annotation of compounds helps a lot, caution is advised regarding the linkage between compounds and bioactivity. An obvious link between a known compound class and an activity does not exclude that other unknown compound classes can contribute to the same type of bioactivity as well. This problem can be addressed by fractionation of crude extracts following bioactivity assays for single fractions as discussed in Chapter 4.4. With respect to this, another pitfall may be the presence of formerly unknown derivatives of a known compound class. Some strain may

produce new derivatives which have not been characterized in previous studies. Hence, a detailed evaluation whether a putatively new compound is just a derivative of a known scaffold is indispensable.

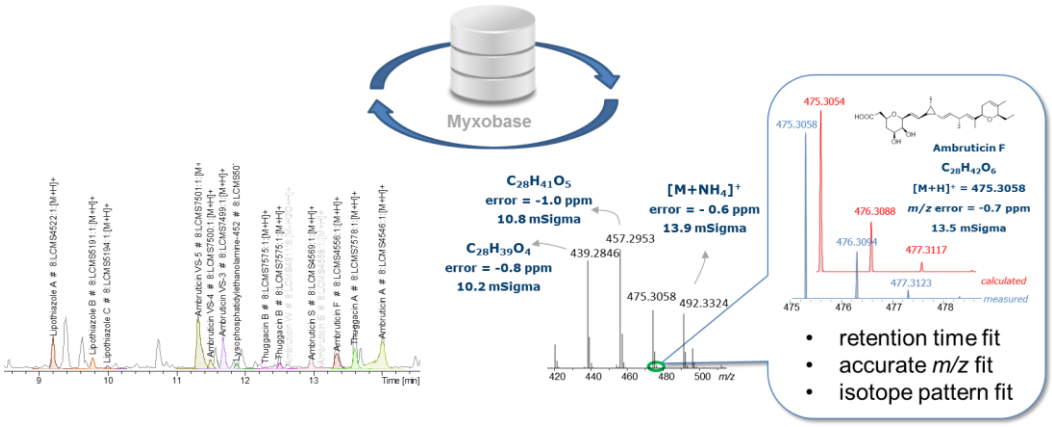


Figure 6: Automatically annotated LC-MS chromatogram (Bruker maXis platform) using the TargetAnalysis™ software. Known compounds are highlighted by color and name tag (left part). Annotation of peaks is based on retention time, m/z, and isotope pattern fit (right part). All data are saved within the Mxbase, our proprietary data base.

In this study the myxobacterial strains *S. cellulosum* So ceGT47 and *S. cellulosum* So ce38 were tested for the presence of new bioactive compounds. At the time when this work was initiated the known secondary metabolites of both strains were limited to merely a handful compound classes as shown in Table 3. As a result of genome mining approaches conducted for both strains it became evident that there is sufficient biosynthetic potential to produce additional compounds. First of all, a brief overview on what is known is given to provide insight into which bioactivities can be expected. Found in So ceGT47, the ambruticins are potent antifungal compounds initially isolated from strain *S. cellulosum* So ce10.[16] Their remarkable structural feature is a cyclopropane moiety in the center of the molecule's chain. The thuggacins are cyclic PKS compounds found in two different myxobacterial genera, *Sorangium* and *Chondromyces*, whereas the derivatives of both genera do only vary in an alkyl side chain.[17,18] The microsclerodermin class was the only known metabolite in So ce38.

In the course of this thesis, the complete structure with all stereogenic centers as well as the full biosynthesis of microsclerodermin was elucidated (Chapter 3). Furthermore, pellasoren was identified in So ce38 as well, alongside with the elucidation of its biosynthesis (Chapter 2). Aside of this, there were new compound classes identified as secondary metabolites such as SH-373, SH-407 and the lipothiazoles in So ceGT47 (see Chapter 4.5.7 for lipothiazole). The metabolites SH-373 and SH-407 are uncommon δ-amidated glutamates that were recently identified by Stephan Hüttel in a *Chondromyces* sp. turned out to be present in numerous myxobacterial strains. In So ce38 two additional compound classes were identified as secondary metabolites albeit not isolated and fully characterized. Soce38-cpd01 was identified by gene knockout of an unassigned PKS cluster and subsequent search for changes in the resulting extracts using statistical analysis. This compound turned out to be highly unstable hampering

its isolation which is partly attributed to its phosphorylation. Soce38-cpd02 is a large, most likely NRPS-derived compound class as judged by its fragmentation pattern. It is noteworthy, that producers of microsclerodermin M (Chapter 3) do produce this compound class as well. However, based on MS/MS data, a relation of this structure to microsclerodermin can be ruled out.

Table 3: Overview of secondary metabolites found in the myxobacterial strains So ceGT47 and So ce38.

Strain	Compound class	known before	this work	activity
So ceGT47	Ambruticin	X	X	antifungal (*C. albicans*, *P. anomala*) antibiotic (*S. aureus*)
	Thuggacin	X		antibacterial, cytotoxic
	Lipothiazole		X	-
	SH-373[a]		X	-
	SH-407[a]		X	-
So ce38	Microsclerodermin	X	X	antifungal (*P. anomala*, *C. albicans*)
	Pellasoren		X	cytotoxic (HTC-116)
	Soce38-cpd01[b]		X	antibacterial (*S. aureus*)
	Soce38-cpd02[b]		X	-

[a] identified and characterized by Stephan Hüttel, [b] not purified, structure not elucidated

4.4 Bioactivity Testing

The driving force behind natural product isolation is the identification and characterization of new pharmaceutically or agrochemically relevant compounds. In light of this it is of utmost interest to include a suitable bioactivity screening in the overall workflow. Crude extracts or compounds showing antibiotic effects are ideally identified by bioactivity assays in an early stage of research, e.g. by monitoring growth inhibition. Although these assays do only cover a small set of representative indicator strains the results point toward promising compounds and thereby initiate further work on a bacterial extract (commonly referred to as an activity-guided approach. For this purpose crude extracts of strains are subjected to a panel of indicator organisms covering Gram negative bacteria, Gram positive bacteria, and fungi. Extracts are applied by means of a dilution series with consecutive 1:2 dilutions starting from undiluted (A) to the highest dilution of 1:128 (H). Comparing bacterial extracts of the same strain grown under different cultivation conditions allows detection of changes in activity strength or even the emergence of new activities.

In this work extracts of *S. cellulosum* So ce38 and So ceGT47, both cultivated with different media, were tested in a standardized assay with the results shown in Table 4. Bioactivity of So ceGT47 extracts is largely constant whereas differences are more pronounced for extracts of So ce38. The antifungal effects of So ceGT47 and So ce38 are at least to a certain extend attributed to the known compounds ambruticin and microsclerodermin. The remaining activities were not linked to these strains before, and are thus of particular interest.

Table 4: Results of bioactivity testing of crude extracts. Letters indicate the strength of inhibition ranging from weak (A) to strong (H). Each consecutive letter resembles a 1:2 dilution of the precedent letter. (Assays and analysis performed by Viktoria Schmitt and Jennifer Herrmann).

Strain	Medium	*S. aureus* DSM346	*E. coli* DSM1116	*E. coli* TolC	*C. violaceum* DSM 30191	*P. aeruginosa* DSM50071	*M. phlei* DSM43070	*M. luteus* DSM20030	*C. albicans* DSM1665	*P. anomala* DSM6766	*M. hiemalis* DSM2656
So ceGT47	A	B	0	0	0	0	0	0	F	H	C
So ceGT47	H	B	0	0	0	0	0	B	F	H	C
So ceGT47	M	B	0	0	0	0	0	A	F	H	C
So ceGT47	P	B	0	0	0	0	0	A	F	H	B
So ce38	A	F	0	0	C	A	A	0	D	D	0
So ce38	H	H	A	A	E	B	A	0	E	E	0
So ce38	M	C	0	0	B	0	C	0	F	G	0
blank	H	0	0	0	0	0	0	0	0	0	0
blank	P	H	0	0	0	0	A	F	0	0	0

For So ceGT47, the question what causes the effect against *Staphylococcus aureus* and *Micrococcus luteus* initiated approaches to track down the active compounds. This was done by HPLC-based fractionation following a convenient approach which directly links the fractionation to MS detection. Two HPLC methods for fractionation are used with one being identical to the standard screening method using acidic eluents and the other method being based on a buffered eluent system to overcome decomposition of acid-labile compounds during workup. In addition, the use of two eluent systems causes different retention times and alters the sensitivity toward distinct compounds thereby increasing the informational content. For both methods a 20.5 min HPLC run is fractionated into microtiter plates with 96 wells resulting in 93 wells with 110 µl per well and spanning 0.22 minutes each. The fractions are dried afterwards and tested for activity against a specific microorganism. By applying this method we define chromatographic regions with a direct link between observed activity and the compounds eluting in this region as identified by MS.

A simple example to illustrate bioactivity testing of HPLC fractions is shown for a So ceGT47 extract. Initial tests using crude extract resulted in an activity against *Staphylococcus aureus* irrespective of the culture conditions. The strain So ceGT47 is known to produce ambruticins and thuggacins albeit both are not reported to be active against *S. aureus*. This suggested the presence of a new secondary metabolite being responsible for this activity or a novel derivative exhibiting changed activity. Testing of HPLC fractions eventually revealed ambruticin S and F to be active against *S. aureus* which was not known before (Figure 7 B). Moreover, two additional active fractions were identified which is an important result only obtained upon HPLC-based fractionation. It should be noted that results from the two different eluent systems are not identical which in turn confirms using both methods. In the end, three compound classes add up to the observed activity against *S. aureus* in crude extract.

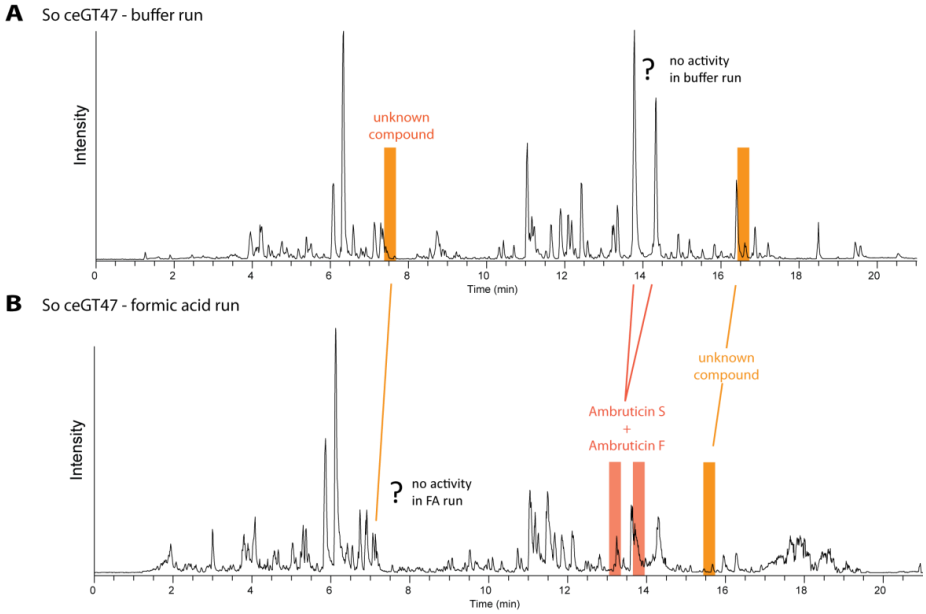


Figure 7: LC-MS chromatograms of So ceGT47 H-medium extract acquired online whilst fraction collection. The same sample was fractionated twice with (A) a 0.1 % formic acid eluent and (B) a 5 mM ammonium formate buffer eluent. Both times, two consecutive runs (V_{inj} = 5 µL) were collected to the same microtiter plate. The plates were dried and tested for activity against S. aureus. Fractionation as well as generation of ions by nano-ESI is done by an Advion Triversa Nanomate. The BPC of such samples is usually a little overloaded which frequently results in bad peak shapes and strong ion suppression (B).

In a similar fashion, fractionation avoids drawing wrong conclusions in terms of an observed bioactivity, like it is the case for ambruticins in So ceGT47. The ambruticins are known to be very potent antifungal natural products. The antifungal activity of the crude extract is easily attributed to this compound class without considering that there may be other antifungal compounds as well (Table 4). Upon fractionation we realized that there is another compound class present with activity against *C. albicans* and *P. anomala* (Figure 8).

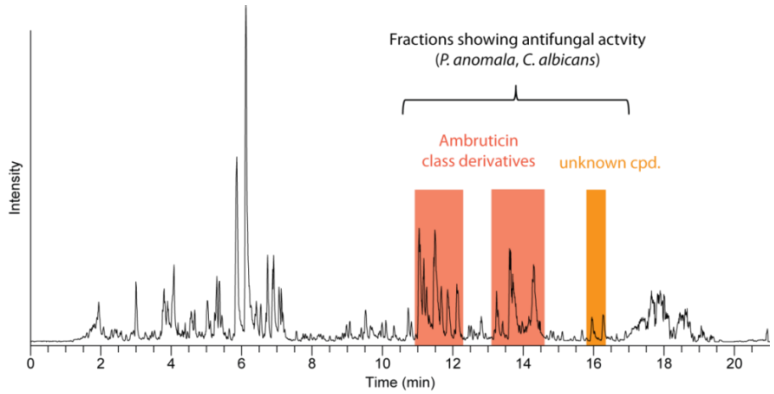


Figure 8: LC-MS chromatogram of So ceGT47 M-medium extract, acquired online whilst fraction collection. The high antifungal activity of the crude extract is not solely attributed to the ambruticin class. An unknown compound class (2 derivatives) contributes to the antifungal activity as well.

These two examples show how a combination of targeted secondary metabolomics and bioactivity screening helps to track down known compounds as well as active unknown compounds.

4.5 Targeted MS/MS in an Untargeted Secondary Metabolomics Workflow

With respect to a screening workflow, the preceding steps yielded very important results by the identification of *known* compounds alongside with the assignment of chromatographic regions exhibiting *bioactivity*. The next step of a screening workflow addresses the search for all *unknown* compounds of which a subset is responsible for the newly observed bioactivity found in some fractions. Using statistical methods is regarded a suitable way to carry out such a task. While various statistical methods have found widespread use in typical metabolomics studies, both data mining in untargeted secondary metabolomics and computational methods for automated analysis of MS/MS spectra are currently still at an early stage. This is partly owing to the fact that comprehensive metabolomics studies lag behind targeted approaches focusing on identification and quantification of a small number of compounds. Methods of untargeted metabolomics or even methods that use MS/MS spectra for creating de novo structural information are fairly new to the rapidly developing field of metabolomics.

Several recently emerged methods use fragmentation spectra from metabolites in order to classify compounds by clustering according to spectral similarity or even to partially elucidate structures, e.g. in case of peptide moieties on the basis of characteristic backbone fragmentation or neutral losses.[5,9,10,19] The focus of this work is to improve the MS/MS data generation step preceding all computational analysis approaches, aiming at “reasonable” data dependent MS/MS acquisition in a sense that high MS/MS coverage of putatively interesting features should be achieved (as opposed to automatically fragmenting the most abundant signals).

The basic assumption of the approach taken here is that bacterial metabolism has a measurable and significant effect on the set of observed LC-MS features (defined by retention time, *m/z*, and intensity). In particular, the differences between feature sets observed from measurement of a “blank” cultivation medium and the same medium inoculated with a bacterium are considered to correlate with growth of the bacterium. Several effects contribute to these differences:

1) Media components are metabolized by growing bacteria
2) Nutrients composed of complex macromolecules are partly digested by excreted enzymes which eventually results in metabolite-like small molecules
3) Cell apoptosis adds additional compounds
4) Complex media can degrade when kept at elevated temperature for prolonged time

Secondary metabolites produced by the bacterium in the course of cultivation are also part of this complex mixture. Thus, analyzing the LC-MS features which newly appear during bacterial growth enables us to direct MS/MS measurements toward the subset of features containing the signals which potentially represent secondary metabolites. A straightforward way to perform comparative analysis in

order to reveal growth-related features is based on measuring two sample types, “strain” and “blank” samples. A “strain” sample is derived from a myxobacterial strain cultivated in liquid medium supplemented with adsorber resin. A “blank” sample is the same liquid medium incubated with adsorber resin but without bacterial inoculate. Both “cultures” are subjected to identical treatment throughout the experiment starting from cultivation, extraction, measurement, up to data processing. The key processing step to prepare LC-MS data for statistical mining is a software-aided bucketing across all samples based on the extracted molecular features (retention time, *m/z*, intensity). More precisely, the presence or absence of a distinct feature is evaluated across all samples, as judged on the basis of retention time and *m/z* pairs. Upon detection of a feature, its intensity value is included into the bucket

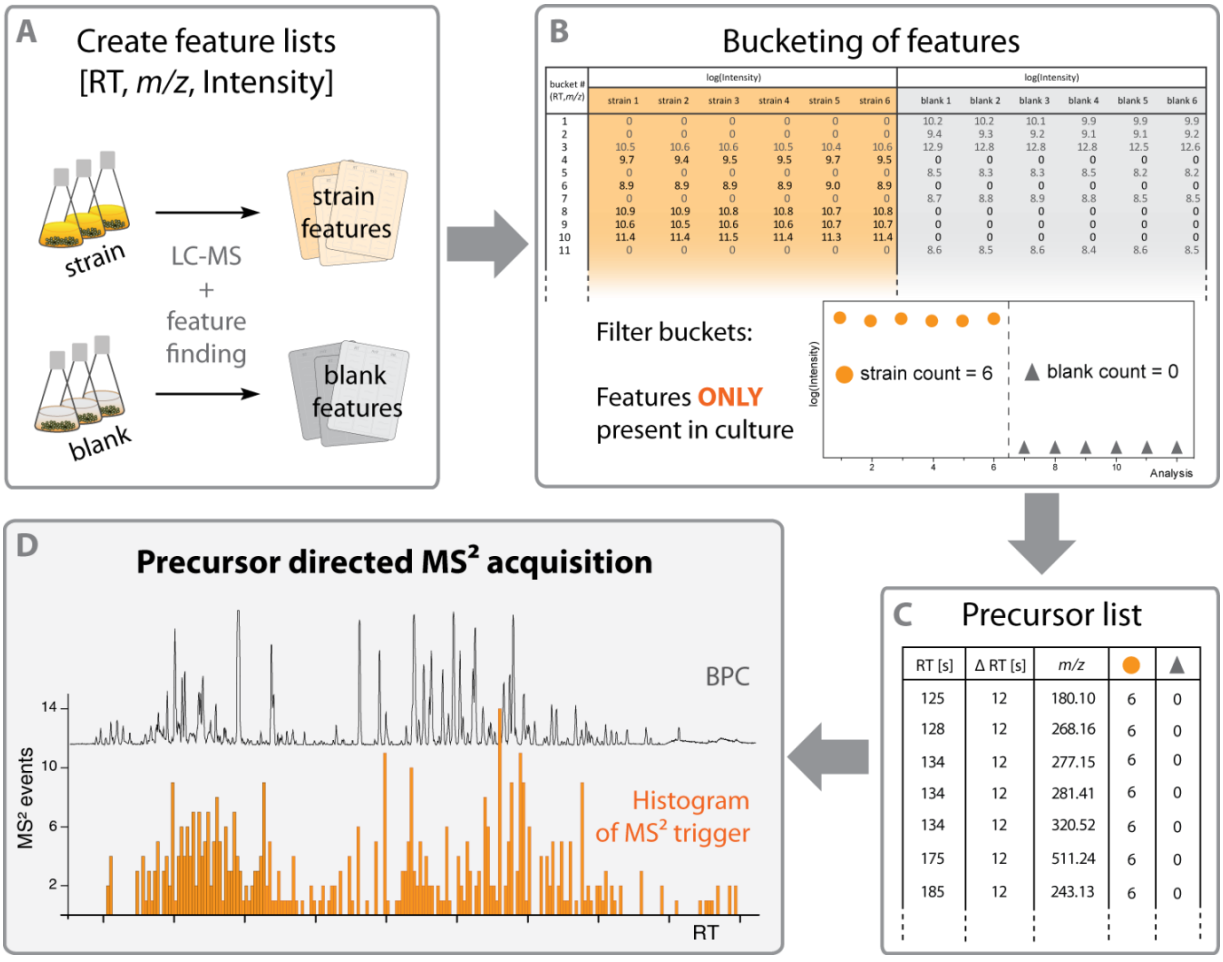


Figure 9: Overview scheme of a targeted MS/MS approach. After creating a suitable sample set the LC-MS data are processed with a feature finding algorithm and transformed to a bucket table covering all features. Filtering this bucket table according to user-specific constraints yields a precursor list that is used for re-measuring the strain-derived extract. MS² events are distributed according to the occurrence of growth-derived compounds in the strain extract.

table. This procedure yields a table covering intensity values of all extracted molecular features for all samples (Figure 9 B). The resulting feature table is filtered according to the constraint that only features found in “strain” samples remain. This workflow results in a list that comprises retention time and *m/z*

for each feature that is related to bacterial metabolism. In a last step, this list is used to specifically target these features as precursors for MS/MS analysis (Figure 9).

Generally spoken, the described processing workflow effectively transforms an intrinsically untargeted secondary metabolomics approach into a semi-targeted method. In the following, details of the established workflow will be explained alongside with a critical evaluation of data generated using this method.

4.5.1 Molecular Feature Annotation

Feature annotation is a crucial prerequisite for any LC-MS-based data mining workflow. A mass spectral feature is a "clean" spectrum representing a single charge state of a single molecular entity, free of noise or other signals. It essentially consists of *m/z* values and the intensities of all related isotope signals. In this study we used the molecular feature annotation algorithm as implemented in the Bruker DataAnalysis software. The "find molecular features" algorithm (FMF) searches for signals that belong together based on their elution profile.

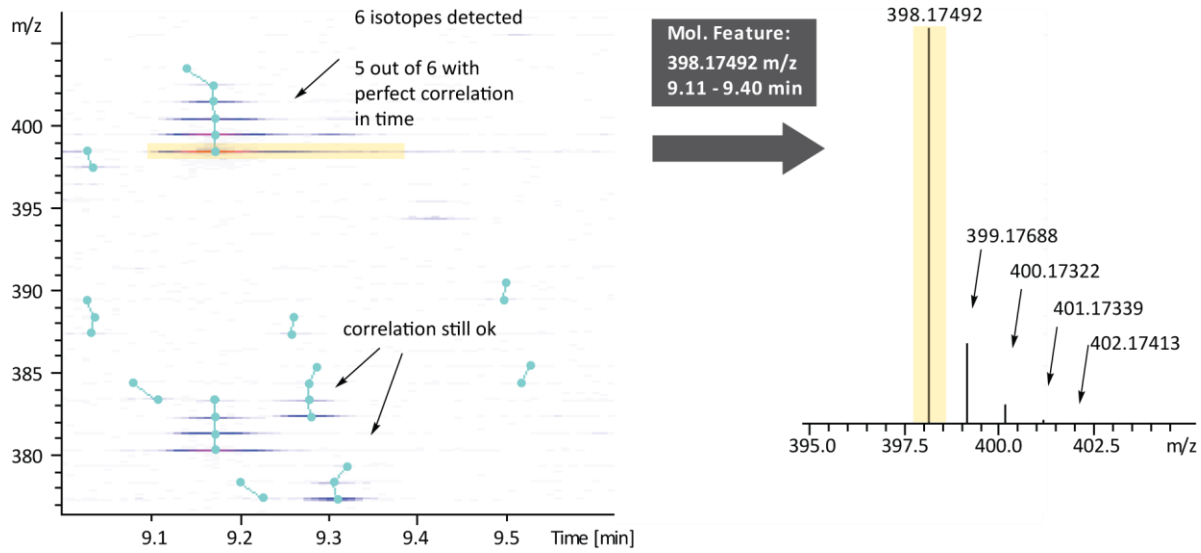


Figure 10: Survey view on LC-MS data highlighting the detection of single isotope (blue dots). Isotope signals belonging to one molecular feature are grouped according to their time correlation as depicted by the blue lines in between the blue dots. The yellow area marks the molecular feature based on the m/z of the first isotope and its time range. A molecular feature is a combination of m/z and intensity values for all isotope signals as well as the retention time range covered by the most intense isotope signal.

More precisely, the algorithm creates extracted ion chromatograms (EICs) of each *m/z* value that is part of a presumed isotope cluster. EIC traces originating from one specific molecular feature reveal themselves through their matching retention time profiles, a finding which is expressed as a correlation factor. Traces that fall into a specific correlation factor range are combined to create a mass spectral feature which comprises the corresponding isotope signals (Figure 10). Different charge states for the same compound are detected similarly and are merged into one molecular feature, thereby reducing

the overall number of mass spectral features. A high feature count appears useful to cover as much information from the LC-MS data set as possible; however, we aimed to restrict the features to "true" chromatographic peaks. For that reason, only features detectable in at least 10 consecutive spectra with an S/N-value ≥ 10 were considered. The number of molecular features identified depends on the bacterial strain, its growth and on the cultivation medium used, as reflected by values in Table 5. Bacterial extracts show an increase in feature numbers compared to the respective blank medium extracts; basically a consequence of both primary metabolism of the growing bacterium and the formation of secondary metabolites. In this experiment we observed between 3000 and 9000 molecular features spanning a retention time window of 19.5 min (Table 5).

Table 5: Molecular feature count for S. cellulosum extracts of strains So ceGT47 and So ce38 alongside with blank extracts for different media used (N = 3 for each value).

10 spectra/feature S/N ≥ 10	So ceGT47	So ce38	blank
P-medium	6729 ± 101	-	4218 ± 286
M-medium	5479 ± 552	3681 ± 60	3322 ± 131
H-medium	4613 ± 266	4634 ± 100	2302 ± 81
A-medium	5752 ± 398	5484 ± 308	2478 ± 309

The high number of molecular features is owing to the rather tolerant settings used for feature finding. Settings are chosen to preserve as much information as possible as a succeeding statistical data evaluation is not limited by a large number of features. The fact that nothing is initially known about the origin of these features is the reason why only moderate filtering should be applied in this very early stage of data processing.

4.5.2 Binning of Features – The Bucket Table

The feature sets of all samples are then used to prepare an overview matrix covering intensity values of all features (Figure 9 B). This type of processing is basically a two-dimensional binning of molecular features across all samples according to retention time and *m/z* value. Note that intensity values are not binned but transferred as distinct entry to the bucket table. First, buckets are generated using the advanced bucketing method of the ProfileAnalysis software (Bruker Daltonics). For this purpose a bucket is created by setting a region defined by retention time and mass accuracy ranges of 0.25 min and 10 ppm around every feature. If there are several features in one bucket due to overlapping signals, the bucket is split until only one feature remains per bucket. In the second step, features of other samples are checked if they fit to one of the existing buckets. Matching features are added to the existing bucket. In case a feature does not fit to any of the existing buckets a new bucket will be created. This

procedure is repeated for all the samples of the experiment until every feature of every sample is part of a bucket. Finally, bucket intensity values are log-transformed to overcome heteroscedasticity.[20] The term heteroscedasticity describes the effect that measurement uncertainty is not equally distributed for the acquired data, i.e. the statistical variance is usually bigger for low abundant peaks. This difference in variance can have several reasons, e.g. metabolite changes due to irreproducible cultivation issues or systematic deviations in mass spectrometric detection. This effect should always be considered when dealing with wide concentration ranges as present in complex biological samples.

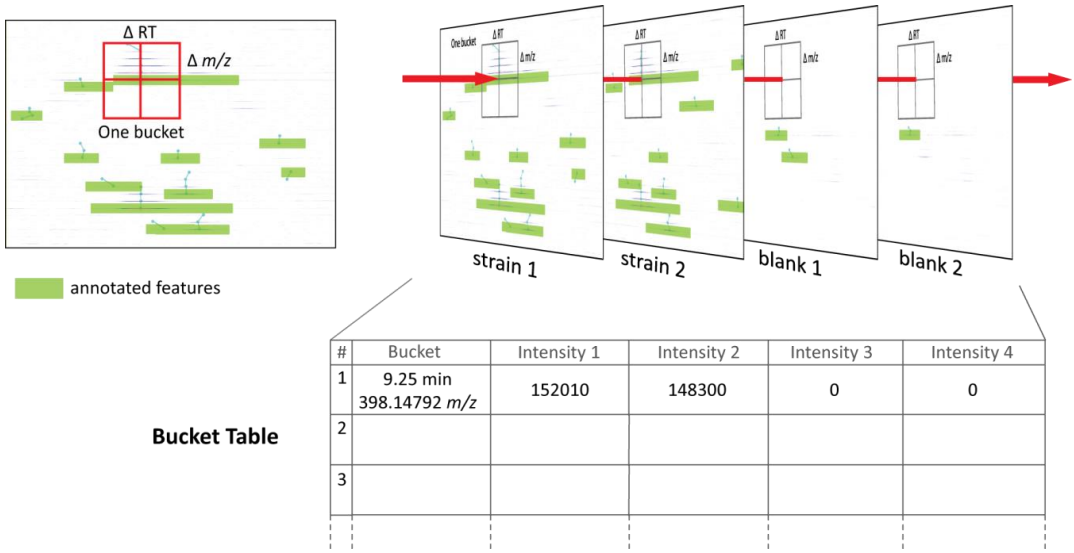


Figure 11: A bucket of predefined RT and m/z range is created around each feature. All samples are checked for features matching the bucket constraints upon which the respective intensity values are saved within the bucket table. The example shows two strain samples having the feature of interest and two blank samples were no features fit to the bucket table.

A bucket table processed by these methods may be used as input for multivariate data analyses. Principal component analysis (PCA) is good technique to visualize the differences between samples and thereby allows a grouping of samples in order to reveal trends and outliers. PCA reduces the dimensionality of a multivariate dataset while the information of the underlying data set is retained. Such a method is of interest when the variations between strain samples and blank samples are of interest (Figure 12). For this type of unsupervised analysis, a Pareto scaling is applied which is commonly used in the metabolomics field as it lowers the impact of highly abundant peaks on statistical model calculation.[20,21]

A PCA scores plot gives a first hint whether the underlying data contains samples which are critical outliers having negative influence on the statistical model. If a single sample is not well explained by the model it has to be removed from further data mining, as it may abnormally affect the final result and lead to invalid conclusion. The expected output of PCA analysis provides the information that is needed for further processing by clearly distinguishing feature sets from “strain” and “blank” samples. The underlying reason for this differentiation is depicted in the loadings plot in which all buckets are shown

(Figure 12). The more off-center a bucket is located, the more this bucket contributes to the observed grouping in the PCA model. A grouping of buckets is frequently observed as a consequence of similar

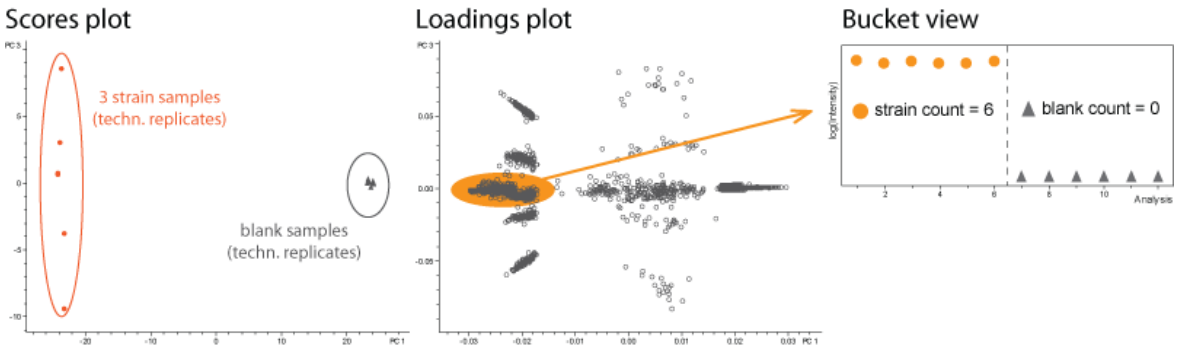


Figure 12: Example of a PCA analysis for a S. cellulosum So ceGT47 sample set highlighting the ability to distinguish between different samples (left part) by grouping the buckets belonging to the data set (right part).

characteristics with respect to the whole data set as depicted with the marked area in Figure 12. This group represents all buckets that are solely found in six out of six "strain" samples but never in "blank" samples (strain count = 6) and thereby represent only compounds that are related to bacterial growth (Figure 9 B or Figure 12). A suitable way to get a full list of such buckets is filtering the respective bucket table according to the given constraints whereas determining reasonable constraints for the underlying data set was one objective of this work. When comparing a set of "strain" samples with the corresponding set of "blank" (medium) samples the question arises how many biological and technical replicates are needed to reduce the number of false positive hits, e.g. originating from inconsistent cultivation conditions or technical issues. For this reason it is not advisable to take solely two data files into account as it will certainly yield vague results, i.e. results based on too many false positive hits witch causes questionable conclusions. An adequate compromise of workload and expected outcome is necessary when preparing samples. To test the effect of technical and biological replicates we used a sample set of *S. cellulosum* So ceGT47 in P medium, consisting of twelve extracts comprising six "strain" extracts and six "blank" extracts where each sample was measured twice. Bucket tables were generated using different numbers of files starting from 2 vs. 2 up to 6 vs. 6 biological samples. Having replicate measurements of all biological samples allowed us to create bucket tables of type 2R vs. 2R up to 6R vs. 6R where "R" indicates the use of technical replicates. All bucket tables were filtered for the strictest condition which is

"bucket count strain"/"bucket count blank" = max value / 0,

e.g. for a 3R-3R data set only buckets of type 6/0 are counted. The amount of buckets for each table is plotted against the number of biological samples to evaluate the effect of the sample set size (Figure 13).

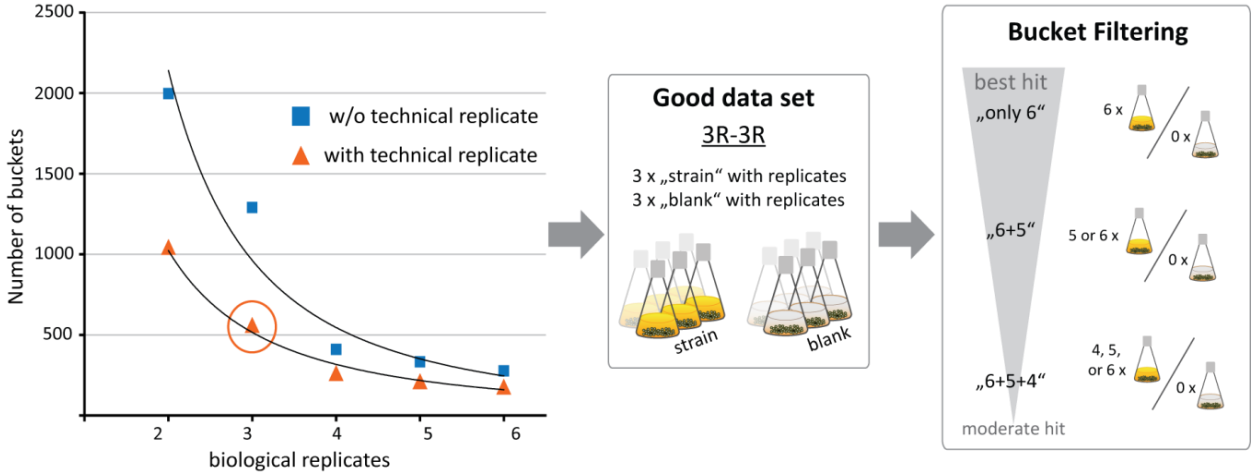


Figure 13: Bucket count in relation the number of samples involved in analysis. The bucket count seems to approach a basal level as a consequence of reducing false-positive hits when using larger sample sets. The benefit of replicate measurements is remarkable when using two or three samples, respectively. A 3R-3R data set is regarded as suitable for data mining.

The decrease in buckets is not linear but rather approaches a basal level. This is expected as noise features are more effectively eliminated when using larger sample sets. An important result is the delineation of impact of replicate measurements upon which we decided to continue this study with data sets that consist of three samples per set with each sample measured twice. Hence, those buckets which are present in every measured sample fulfill the condition “strain”/”blank” = 6/0. We also consider buckets that match a 5/0 and 4/0 condition, based on practical considerations since the 5/0 condition can mitigate the influence of one failed measurement e.g. if it faced technical problems. A 4/0 condition even allows for additional drop-out of a biological sample, e.g. owing to inappropriate growth of the culture.

4.5.3 Creating a MS/MS Precursor List

The filtered bucket table is the basis for a precursor list that is used in subsequent LC-MS/MS measurements. It is important to note that bucket tables exported from Bruker ProfileAnalysis 2.1 (build 250) are not suitable to use them directly as a precursor list albeit they basically comprise RT and *m/z* values. However, each bucket is labelled with the *m/z* value of the respective $[M+H]^+$ ion even if only higher charge states were detected, i.e. a compound only showing up as +2 ion will get a bucket label that refers to the +1 ion *m/z* value.

As a technical workaround, charge state information can be preserved by exporting distinct lists (RT, *m/z*) for each charge state. After creating charge state lists for +1, +2, +3, +4, and +5, each row of the charge state list (RT, *m/z*) is searched amongst the complete bucket table using a SQL query (Figure 14). Upon a matching pair, the respective row in the bucket table is extended by the charge state information. This downstream processing is realized by KNIME software (Konstanz Information Miner software) in combination with a MySQL database using a workflow that annotates observed charge

states to each bucket. The SQL query was repeated for all rows of all charge state lists to end up with a bucket table covering charge state information for every bucket. It is moreover necessary to allow for some RT and *m/z* deviation when running the SQL query as the automatically exported charge state files do not exactly fit to the bucket labels, i.e. entries in the charge state list comprise time windows while the bucket table has fixed values. As there is no way to circumvent this, all time windows were replaced by the mean values and these mean values compared to the bucket table by allowing a narrow retention time tolerance. This processing yields a precursor list with full charge state information which is further used to calculate the *m/z* value based on each charge state.

```
SELECT
        buckettable.RT_bucket, buckettable.mz_bucket,
        chargelistX.time, chargelistX.MZ_SPL
FROM
        buckettable, chargelistX
WHERE
        (chargelistX.time BETWEEN (buckettable.RT_bucket-0.2)
        AND (buckettable.RT_bucket+0.2))
AND
        (chargelistX.MZ_SPL BETWEEN (buckettable.mz_bucket-0.1)
        AND (buckettable.mz_bucket+0.1))
```

Figure 14: SQL query for identification of charge states of each bucket table row. The query is performed for each charge state list (X = 2, 3, 4, 5) and the query result is added to the bucket table for each list.

After applying an intensity threshold, all buckets with intensity values less than 10,000 are excluded from the list. This bucket table is again filtered according the “strain count”/”blank count” constraints and exported from KNIME as three independent lists (RT, ΔRT, *m/z*) named “only-6”, “6+5”, and “6+5+4” referring to the “strain” count that is used in each list (Figure 13). This type of list covers all the information necessary to perform a precursor selection for MS/MS analysis. All molecular features that are most likely related to bacterial metabolism – as based on statistical data evaluation – are part of this list. The size of such a scheduled precursor lists (SPL) spans from 463 to 922 entries for a 19.5 min chromatographic run of So ceGT47 extract cultivated in P medium. A histogram view of the “SPL_only-6” presented in Figure 15 illustrates in which time slot acquisition of MS/MS spectra will be triggered.

With this precursor lists in hand we set out to measure samples in SPL-MS² mode and in auto-MS² mode. Noteworthy, the sample material used for MS/MS analysis is always a mixture of the strain extracts that were used to create the bucket table. This enables analysis of “4/0” bucket scenarios as those may be related to a feature physically missing in one of the three samples. Only a mixture of the three samples can cover all “4/0”-related features.

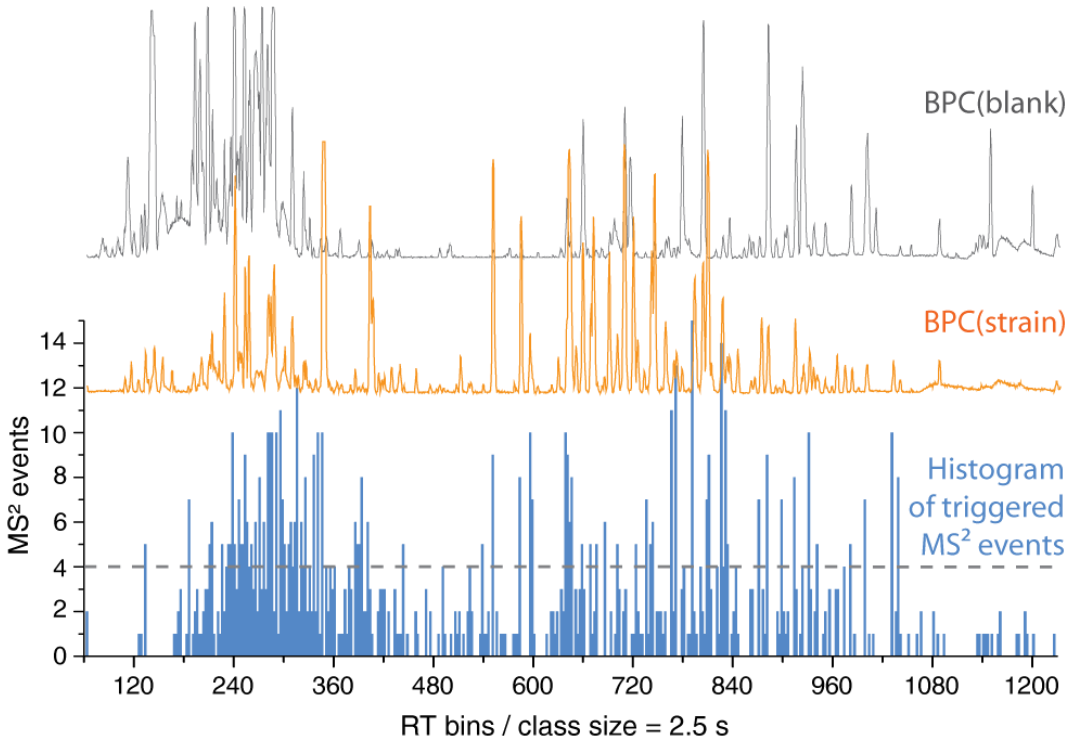


Figure 15: Overlay display of base peak chromatograms of a P-medium blank extract (grey) and a So ceGT47 P-medium extract (orange). The blue graph is a time-based histogram of the MS² events that are listed in one of the SPL used for LC-MS/MS. The number of MS² events resembles the orange chromatogram.

4.5.4 MS/MS Data Acquisition and Processing

The next step of this data mining workflow is the acquisition of MS/MS spectra using the trigger events as defined in the SPL which was generated in the preceding steps. In particular, the mass spectrometer is supposed to use the information on retention time and m/z value as saved in the SPL to exclusively select matching ions as precursors for subsequent MS/MS analysis. Ideally, the instrument should be able to select all desired precursors and perform the requested high-resolution MS/MS measurements. However, full coverage of the SPL may not be achievable for several reasons:

1) Compounds responsible for low abundant peaks may circumvent detection in subsequent MS/MS measurements upon degradation
2) Some SPL entries are false positives owing to errors during feature bucketing
3) Time slots may cover more MS/MS triggers than the instrument duty cycle can cope with

Especially the latter can be addressed by optimization of the MS/MS parameters and considering some practical constraints. As a result of small chromatographic peak widths of only 6 to 10 seconds and the complexity of the sample, SPL entries are dense with respect to the retention time scale. This is exemplified by a histogram for a typical SPL with around 500 entries revealing time slots with frequently more than 4 precursors (Figure 15). Especially time slots with a high number of potential precursor ions could be a problem for the overall coverage of SPL entries. This is owing to a prolonged duty cycle which

eventually causes a subsequent time slot not to be taken into account. In order to avoid interference with subsequent time slots, the duty cycle time should be limited, e.g. by restricting the number of MS/MS scans per full scan. If a MS^2 method is used which is limited to a maximum of 4 precursors (dashed line in Figure 15) some of the listed precursors for this particular time slot will be missed for the sake of a better coverage of all time slots. However, one of the methods tested has a strict precursor selection which causes all precursors listed for a given time slot being addressed, no matter how many are affected.

The origin of such time slots loaded with SPL entries is partly attributed to compounds susceptible to strong in-source fragmentation as respective fragment ions are individually added to the potential precursor list during the initial statistical approach. In principle, applying MS/MS fragmentation to those in-source derived fragments can increase confidence for compound identification, but on the other hand these data add redundancy to the dataset. Hence, restricting the number of precursors of a duty cycle can help to increase coverage for diverse compounds alongside with avoiding redundant information. With respect to this problem, an automatic adduct finder algorithm could help to combine frequently observed adducts and neutral-loss-derived features into a single feature.

In summary, a compromise between cycle time, spectra quality, and number of selected precursor ions had to be evaluated as part of this project. Several MS/MS settings were tested to check for a high SPL coverage. In particular, the effect on SPL coverage was tested with settings varying in the type of precursor selection (auto-MS^2 or SPL-MS^2), the number of precursor ions, and scan speed. The duration of each duty cycle is a function of the respective precursor intensity and the time necessary for one full scan and all preceding MS/MS scans. It is noteworthy, that for SPL-MS^2 methods the number of precursors is variable but limited to a maximum. Hence, duty cycle time in SPL-MS^2 acquisition can vary depending on both, the number of precursors for a given time slot and the precursor intensity. On the contrary, duty cycles for auto-MS^2 methods do only vary based on precursor intensity for the methods used in this work.

The most important value to rate the quality of these analyses is the “SPL coverage” which reflects how many SPL entries were fragmented during one run. To achieve coverage of 100 %, each SPL entry needs to be fragmented at least once. Since the SPL coverage is not affected when a single precursor is measured several times, an additional value named “MS/MS efficiency” is introduced. This value reflects the percentage of MS/MS events that act on precursors which are listed in the SPL. MS/MS efficiency is generally not 100 % even though a SPL-guided precursor selection is used. This is owing to the fact that precursor selection constraints were not that strict, i.e. a precursor *m/z* was selected if the determined value matched within a window of +/- 0.2 *m/z* to the value as listed in the SPL. When using complex biological samples there is a chance that some impurities match the constraints as well. In light of this it becomes clear that a 100 % value is usually not achievable albeit it would reflect the ideal situation.

Table 6: Different MS/MS methods as used in this study. Changes are related to acquisition speed, number of precursors and the way precursors are selected, i.e. automatically based on most abundant ions (auto-MS²) or based on pre-filtered precursor lists (SPL, scheduled precursor list). Measurements were done in duplicate for each setting.

#	mode	precursor	acq.speed #1 [Hz]	acq.speed #2 [Hz]	Prec. Int. #1	Prec Int. #2	duty cycle [s]
1	auto MS²	2	1.5	2.9	1E+04	5E+04	1.2 - 1.8
2	auto MS2²	3	1.5	2.9	1E+04	5E+04	1.5 - 2.4
3	auto MS²_fast	3	1.5	6.7	1E+04	1E+05	0.9 - 2.4
4	SPL 2prec	2	1.5	2.9	1E+04	5E+04	1.2 - 1.8
5	SPL 3prec	3	1.5	2.9	1E+04	5E+04	1.5 - 2.4
6	SPL 3prec_fast	3	2.0	6.7	1E+04	1E+05	0.9 - 2.0
7	SPL 4prec_fast	4	2.0	6.7	1E+04	1E+05	1.1 - 2.5
8	SPL strict	all	2.0	10.0	1E+04	1E+05	-

In conclusion, “MS/MS efficiency” and “SPL coverage” are both used to rate the methods whereas “SPL coverage” is of higher explanatory value. Aside of this, it is of interest to analyze how many known compounds – as identified by database query – are captured by the MS² events of one measurement. To address this question, reports of known compounds were created using the Bruker TargetAnalysis software and manually curated. These reports were compiled for each strain/medium combination in this study and used to search for matching MS/MS events in an analysis. The processing is accomplished by a combination of Bruker DataAnalysis, KNIME, and a MySQL database. First, MS/MS events of a given measurement are exported from DataAnalysis and imported to the MySQL database covering the necessary information, i.e. retention time, *m/z*, and intensity of each precursor ion. In the next step the SQL language is used to compare the SPL with the list of MS/MS events using the query shown in Figure 16. Matches must fulfill the constraints of a 0.4 min retention time window and a 0.04 *m/z* window around the distinct SPL values.

```
SELECT
        SPLlist.RT_SPL, SPLlist.MZ_SPL, SPLlist.Intensity, SPLlist.charge,
        MS2event.RT_exp, MS2event.MZ_exp, MS2event.Intensity, MS2event.I_fragments

FROM SPLlist, MS2event

WHERE
(MS2event.RT_exp BETWEEN (SPLlist.RT_SPL-0.2) AND (SPLlist.RT_SPL+0.2))
AND
(MS2event.MZ_exp BETWEEN (SPLlist.MZ_SPL-0.02) AND (SPLlist.MZ_SPL+0.02))
```

Figure 16: SQL query used to identify SPL entries being measured in a real sample.

MS/MS event lists were compared to the corresponding SPL, e.g. a SPL-MS² sample measured using a list of type “SPL-only6” is compared to the same list by means of a SQL query. In line with this, auto-MS² samples are compared to lists of different type such as “SPL-only6”, “SPL-6+5”, and “SPL-6+5+4”. In

addition, all MS/MS events were compared to the corresponding list of known compounds derived from the TargetAnalysis report using a similar query (Figure 17).

```
SELECT
        known.RT_TA, known.MZ_TA, known.`Name`, MS2event.RT_exp, MS2event.MZ_exp,
        MS2event.Intensity, MS2event.I_fragments

FROM    known, MS2event

WHERE
(MS2event.RT_exp BETWEEN (known.RT_TA-0.2) AND (known.RT_TA+0.2))
AND
(MS2event.MZ_exp BETWEEN (known.MZ_TA-0.02) AND (known.MZ_TA+0.02))
```

Figure 17: SQL query used to identify the number of MS/MS scans related to known compounds within list of MS/MS events.

Although detailed information on matching entries is available, the classification of the methods used here considers solely the sum of all hits. For example, 80 % SPL coverage is achieved if 400 out 500 SPL entries were fragmented irrespective of the detailed characteristics underlying each entry, e.g. whether the spectrum shows a lot of fragment signals or whether the precursor was fragmented once or twice.

The effect of different MS/MS acquisition methods is depicted in Figure 18. First of all, an obvious difference between auto-MS² and SPL-MS² methods is observed with respect to both SPL coverage and MS/MS efficiency. Effects within each type of method are less pronounced as for example auto-MS² methods show a slightly higher SPL coverage when using 3 instead of 2 precursors (#1 to #3, Figure 18). At the same time MS/MS efficiency is lowered indicating that more fragment spectra were executed although those are not related to the potentially interesting SPL. In terms of SPL-MS² methods, speeding

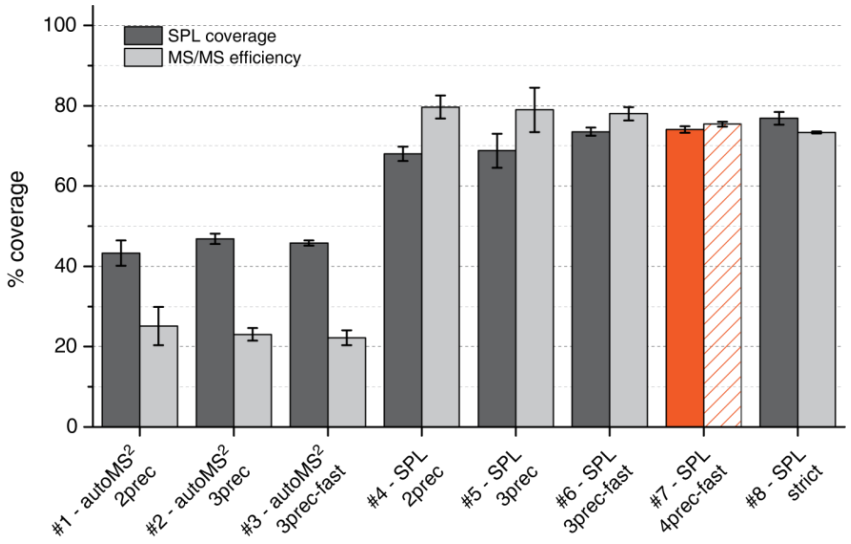


Figure 18: Effect of different MS/MS settings on SPL coverage and overall MS/MS efficiency (N = 2 for all values). An SPL coverage of 100 % is equal to acquisition of at least one MS/MS scan for all precursors listed in the SPL (463 entries for this list). The MS/MS efficiency is the relative number of single MS/MS scans that were related to an SPL entry. The bars highlighted in orange represent the setting that is used for all upcoming measurements.

up the duty cycle time has a weak effect on SPL coverage, e.g. if comparing setting #5 and #6. However, SPL coverage increases when 4 precursors are taken into account while there is no change between 2 or 3 precursors (settings #4 to #8). This is reasoned by the higher coverage in time slots with many SPL triggers. Expectedly, the best coverage is achieved for the strict precursor selection when all precursors that are listed are actually subjected to fragmentation. However, this setting can lead to the above mentioned problem of prolonged duty cycles. Using 4 precursors seems to be a good compromise between coverage and scan speed.

Taken together, the devised MS/MS method in this study was set to allow a maximum of 4 precursors using a full scan acquisition time of 2 Hz followed by MS/MS spectra acquisition at variable scan speed (method #7, Table 6). In detail, scan speed in MS/MS scans is governed by precursor intensity and follows a linear increase from 2 to 6.7 Hz between the two intensity set points yielding duty cycles ranging from 1.1 to 2.5 seconds. The mass spectrometer's software reads the SPL for the upcoming 2 seconds and fragments the four most intense precursors in this time slot. After this duty cycle the upcoming 2-second-window is evaluated. Each precursor is moved to an exclusion list when 2 fragment spectra were acquired and remains there for 0.2 minutes. Note that auto-MS² mode selects the 3 most abundant *m/z* of a full scan and fragments those using the same scan speed settings as for the SPL mode. Hence, the only difference in these experiments is in fact the approach taken for precursor selection.

Apart from optimizing acquisition settings, full SPL coverage is hard to achieve since some components of the crude extract are susceptible to degradation. Indeed, in-depth manual analysis reveals that missing SPL precursors most often show up with low intensities in the bucket table, e.g. an intensity median value of 19,000 is given for the 100 missing SPL entries in case of the above shown strict method (method #8, Table 6). Thus, these precursors were already low abundant at the time the extracts were initially measured and may escape MS/MS fragmentation in a subsequent run. Stability drawbacks can be overcome when all samples are measured in a shorter time, which is indeed feasible once a suitable method has been established.

4.5.5 Evaluation of SPL-derived Data

A first impression of the usefulness of a SPL-MS² method has been given in the preceding chapter when the precursor selection of auto-MS² is compared to SPL-MS². However, a more detailed evaluation is necessary to estimate the efficacy of the whole approach. The initial feature annotation and the following statistics-based bucketing (Chapter 4.5.1 and 4.5.2) may already introduce false positive entries to the SPL. As bucketing is – like every technique of such type – error-prone, there are likely features that were assigned to the strain group although they were present in blank samples as well. Such entries should be detectable by applying the respective SPL-MS² method to a medium blank

sample with the results shown in Table 7. Here, up to 5 % of the SPL list entries were fragmented in a blank sample run and most part of this can be attributed to wrong SPL entries.

Table 7: Comparison of blank and strain samples when measured with a SPL-only6 method. The table shows the number of MS/MS events measured in each run together with all MS/MS events that are related to a SPL entry, no matter if scans belong to identical precursors. The SPL coverage is calculated based on filtered MS/MS scans to account for scans that belong to identical precursors (not shown in this table). Values are average values of duplicate measurements.

Sample	all MS/MS events measured	MS/MS events related to SPL entries	SPL size	SPL coverage
blank, P medium	56	23	459	**4%**
blank, A medium	66	37	483	**2%**
blank, M medium	116	39	485	**5%**
blank, H medium	52	13	436	**2%**
So ceGT47, P medium	683	583	459	**76%**
So ceGT47, A medium	651	514	483	**68%**
So ceGT47, M medium	586	503	485	**70%**
So ceGT47, H medium	647	566	436	**78%**

The table reveals a problem which is related to precursor selection and the constraints used therefor. There are MS/MS scans which are not directly linked to the SPL entries causing the discrepancy between the first two data columns of Table 7, e.g. 56 scans were measured but only 23 are of relevance in case of P-medium blank. These additional MS/MS scans are related to – mostly low abundant – peaks or chemical noise matching the constraints for precursor selection without actually being the correct precursor. This effect is pronounced with increasing complexity, e.g. when comparing the highly complex strain samples to the moderately complex blank samples. Higher complexity does usually come along with more chemical noise and thereby increases the chance to accidently select a wrong precursor. By adjusting the constraints for precursor selection in terms of *m/z* range these wrong selections may be further reduced if necessary. The end, the comparison of blank and strain samples measured with identical SPL-MS² mode clearly shows that some of the SPL entries are wrong as they are also found in blank samples. However, a false positive rate of usually less than 5 % is an acceptable level of confidence.

Another approach to prove the usefulness of SPL-MS² is based on a serial dilution experiment. Two crude extracts were diluted with respective blank extracts in order to maintain or even increase the complexity of the sample while at the same time the concentration of strain-related compounds is decreased. As a consequence, strain-related peaks will be suppressed by matrix peaks from the blank sample or simply too diluted to be detected or selected as a precursor by auto-MS². Each sample was

measured with auto-MS² and with SPL-MS² using a SPL of type "SPL_6+5". For both methods, auto-MS² and SPL-MS², the SPL coverage (dark grey bar, Figure 19) decreases to a comparable extend with the coverage being better for all SPL-MS²-derived runs. The overall decrease of coverage in the course of a dilution series is reasoned by the fact that signals fall below the limit of detection. The notable result of this experiment is that SPL-MS² methods are capable to use MS/MS scans more efficiently. In detail, the number of overall MS/MS events (red line, Figure 19) is always lower for SPL-MS² methods than for auto-MS² methods while at the same time more precursors of potential interest are subjected to MS/MS scans (light grey bars remain high).

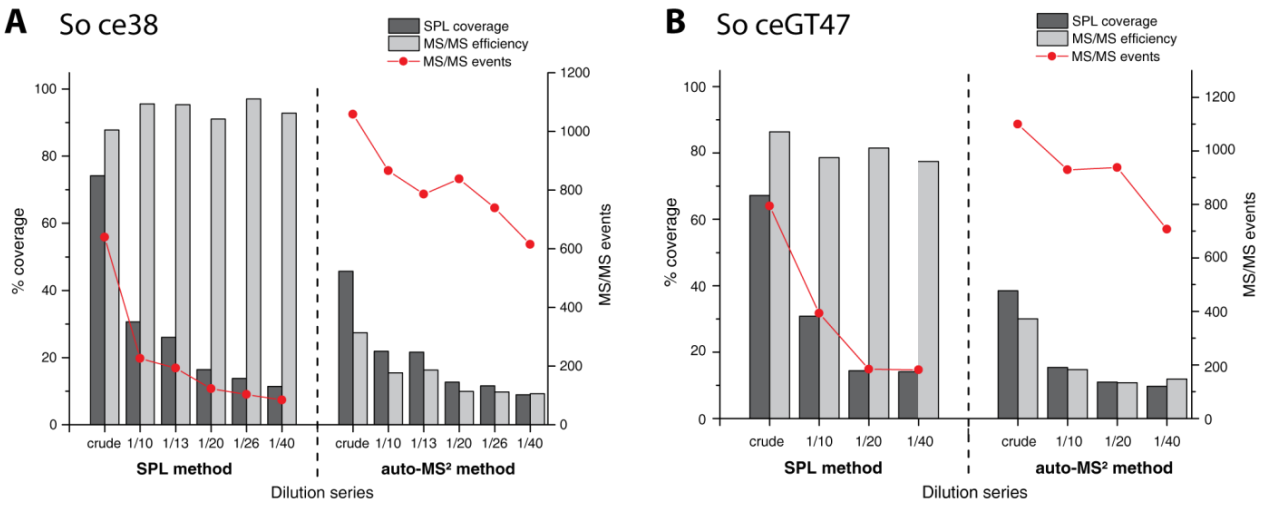


Figure 19: Dilution series of a crude extract of (A) So ce38 – H medium and (B) So ceGT47 – P medium. Left part of each graph shows the SPL coverage when using a SPL-MS² method. Right part of each graph features the same samples measured with an auto-MS² method. The red points represent the number of overall MS/MS scans performed for each setting.

Some of these results are depicted with the help of Euler diagrams which allow a comprehensive view on the efficiency of the method. For this purpose, Euler diagrams were created to represent the overlap between all MS/MS events of a run in comparison to the SPL entries and the list of known compounds as determined for the respective measurement. On the basis of the dilution experiment described before, Figure 20 shows three data acquisition conditions for So ceGT47 P-medium extract diluted with P-medium blank. Note that the figure is based on the same data as used for Figure 19 B. The MS/MS efficiency of SPL-MS² methods becomes clear when the size of the orange circles (MS/MS events) is considered. A dilution causes less MS/MS scans but still a higher coverage of SPL entries for SPL-MS² methods (numbers highlighted in bold font). The right part of Figure 20 reveals that auto-MS² uses a comparable amount of MS/MS scans for each dilution but misses significantly more SPL entries as those are not anymore the dominant peaks of a spectrum owing to dilution.

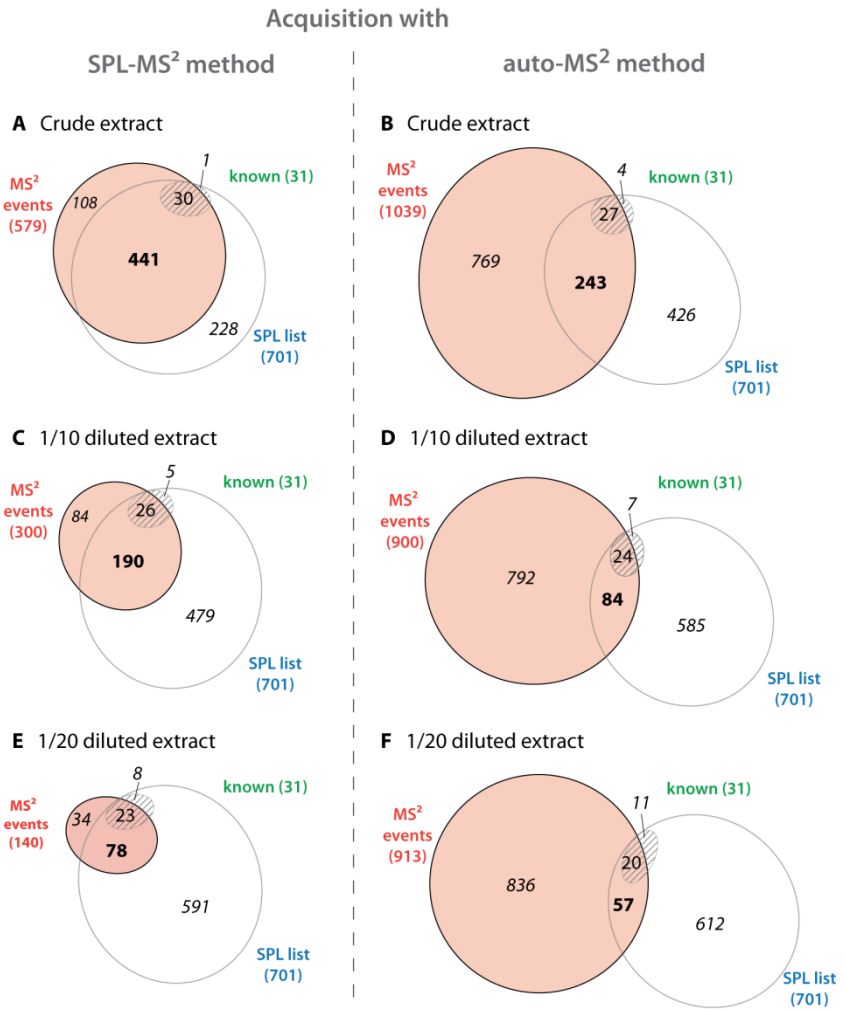


Figure 20: Evaluation of the effect of SPL-based LC-MS/MS acquisition for a dilution series. The data is based on So ceGT47 P-medium extracts which were diluted with P-medium blank in the ratios 1/10 and 1/20 and measured under the control of a SPL-MS² method or an auto-MS², respectively.

Another set of Euler diagrams was created for measurements based on an identical sample measured with different SPL (Figure 21). The idea behind this experiment is to rate the effect of SPL size on SPL coverage, i.e. does SPL coverage remain high even though the SPL size is significantly increased? This would be the case if the system is able to handle the additional MS/MS triggers predetermined by the SPL. This was checked by using three different SPL for So ceGT47 P-medium:

(1) SPL-only6: Buckets found in 6 out of 6 samples (459 SPL entries)
(2) SPL-6+5: Buckets found in 6/6 or 5/6 samples (701 SPL entries, 1.5-fold “only6”)
(3) SPL-6+5+4: Buckets found in 6/6, 5/6, or 4/6 samples (780 SPL entries, 1.7-fold “only6”)

A SPL-MS² run with “SPL-only6” addresses 321 out of 459 SPL entries (70 %). By using “SPL-6+5” and “SPL-6+5+4” the SPL size is increased by factor 1.5 and 1.7 whereas the number of MS/MS scans is increased by a factor of 1.4 and 1.5. At the same time the SPL coverage remains high with around 65 %. This result is depicted by Figure 21 where a high central overlap region indicates the high SPL coverage

throughout the SPL-MS² runs. With respect to this result it is concluded that the significantly increased SPL size is still adequately processed by the instrument. Again, there is a remarkable difference to the auto-MS² methods which addresses only 35 to 40 % of the SPL entries although having more than double the number of MS² events. In terms of the known compounds, no significant difference is observed for both acquisition methods; this result is however anticipated owing to the fairly high abundance of known compounds found in this crude extract.

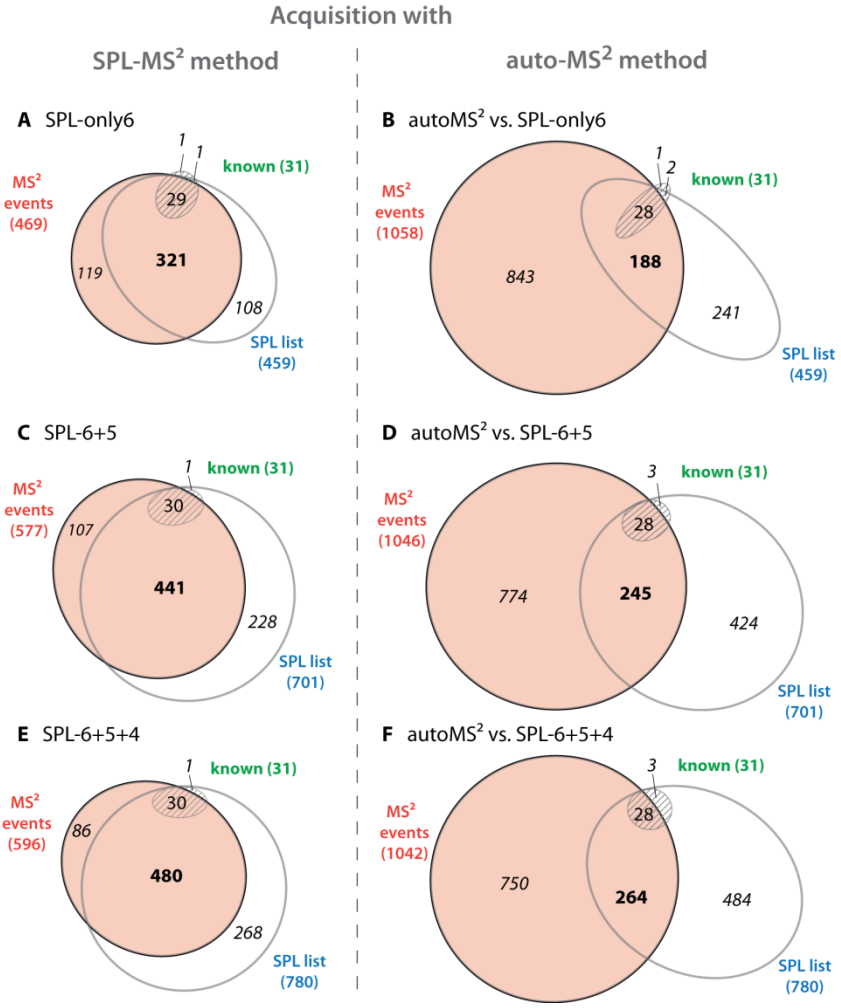


Figure 21: The effect of using different SPL for the same sample represented by Euler diagrams. The overall areas as well as the overlap areas are size-proportional to the values that contribute to the respective field. Data derived from S. cellulosum So ceGT47 P-medium data set.

In a next step, the reason for the discrepancy between methods auto-MS² and SPL-MS² is investigated. In particular, why is the auto-MS² method missing so many of the potentially interesting precursors? Can the initial assumption be confirmed that only the intensity-based precursor selection is causing this effect? A suitable approach to answer this is to check the characteristics of the overlapping regions in the Euler diagram of Figure 21 A and B where one method results in 321 SPL entries measured and the other in only 188. The intensity data for each precursor of these SPL entries was used to create a histogram based on intensity value binning following the creation of an overlay display to directly

compare the SPL-MS²-derived hits with the auto-MS²-derived ones (Figure 22). This was done for a "SPL-only6" and a "SPL-6+5+4" measurement which significantly differ in the overall SPL size as well as in this overlap region.

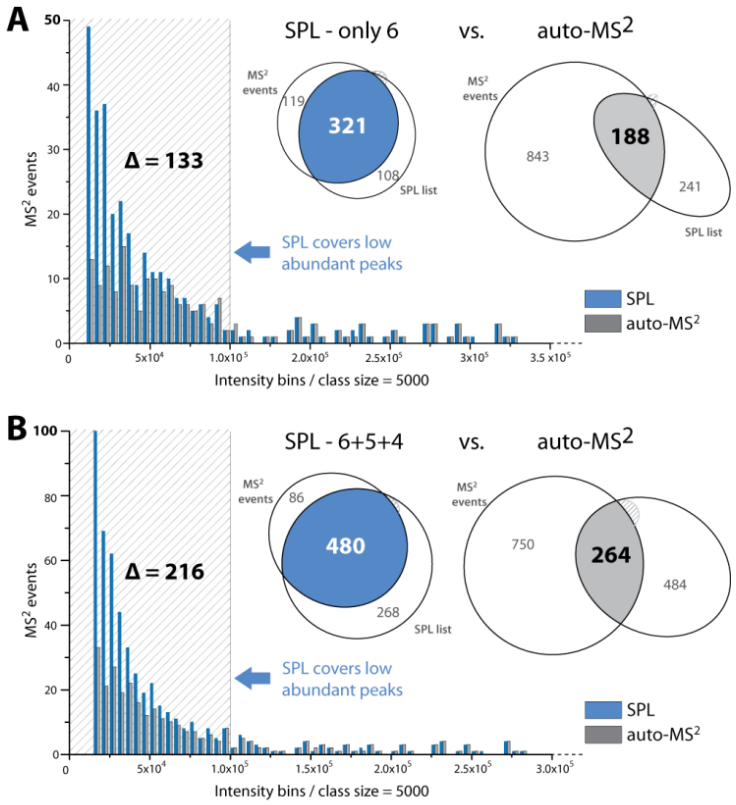


Figure 22: Overlay display of MS/MS event histograms for auto-MS² and SPL-MS² data. The additional MS/MS events of a SPL-MS² run compared to an auto-MS² run are usually found for precursor intensities of less than 1x10^5 counts (highlighted area within the histograms).

The bars reflect the number of MS/MS events with respect to the precursor intensity used in each case. With this graph it becomes evident that using a SPL-MS² method enforces the selection of low abundant precursor ions (blue bars). More precisely, the most part of the additional 133 MS/MS events of the first example of Figure 22 is related to precursor ions with intensities less than 1x10^5 as can be seen by the difference of blue and grey bars in the highlighted area. A similar trend is observed for large precursor lists, i.e. when using the "SPL-6+5+4" as shown in Figure 22 B.

This proves that using a SPL-MS² method enforces the acquisition of precursor ions irrespective of their relative intensity. Furthermore, it was shown that the benefit of a SPL-based MS/MS acquisition in terms of SPL coverage and MS/MS efficiency persists for all sample sets measured in the course of this work, i.e. two different myxobacterial strains using four different media. For all sample sets that were tested, the superior selection of precursor ions was observed when using SPL-methods. In this study, an auto-MS² run covered a maximum of 50 % of the SPL entries which is reasoned by the selection of the abundant peaks of interest. However, the auto-MS² method misses the low abundant peaks of interest

although these should be considered important in terms of being related to bacterial growth (as revealed by the initial statistics-based filtering step)

It should be noted that the capability to select such precursors of lower abundance is only of interest if the resulting fragment spectra bear enough informational content for further use, i.e. by manual evaluation or computational techniques. This requests an evaluation of MS/MS spectra quality when spectra are derived from low intensity precursors. A suitable approach to do so is based on fragmentation of known compounds using the SPL method as established in the course of this work. The concentration of known precursor was reduced on purpose by diluting crude extract with blank extract until a dilution was reached where auto-MS² was not able to select this particular precursor anymore. This diluted extract was then measured using a SPL-MS² method and indeed, relevant precursor ions were still selected for fragmentation. The fragment spectra derived from these measurements were then compared to reference spectra in order to evaluate the spectra quality. This experiment was carried out for three known compounds; two tripeptides named “SH-myxo373” and “SH-myxo407” which are both found in So ceGT47 crude extract as well as for microsclerodermin M which is found in So ce38 extracts. The tripeptides are a new class of δ-amidated glutamates which were recently structurally elucidated by Stephan Hüttel at the Helmholtz Institute for Pharmaceutical Research, Saarland, Germany.

Figure 23 shows the result for compound “SH-myxo373” from the So ceGT47 extract where the $[M+H]^+$ ion is missed in an auto-MS² acquisition but successfully fragmented by SPL-MS². The figure features the full scan spectrum of the auto-MS² and the SPL-MS² method with selected precursor ions marked red (Figure 23 A, B). Moreover, the resulting MS/MS spectrum (Figure 23 C) features 6 out of 7 relevant fragment ions when compared to a reference MS/MS spectrum of this compound (Figure 23 D). The good quality of this fragment spectrum could be reasoned by the moderate intensity of the precursor ion in the range of $5*10^4$ counts. With respect to the LC-MS system used for this work, signals with intensity lower than $1*10^4$ counts frequently result in fragment spectra with less informational content.

Figure 24 is related to the second tripeptide named “SH-myxo407” which is also found in So ceGT47 extracts. For this particular compound the fragment spectrum shows only few signals albeit these are all in agreement with the reference spectrum. Together with the accurate mass determination and isotope pattern fit of the parent ion, this result is sufficient to tentatively classify this compound. The spectra quality is also sufficient to recognize that both peptides “SH-myxo373” and “SH-myxo408” belong to the same class, thus the fragment data fits the needs for manual evaluation and likely also computational techniques.

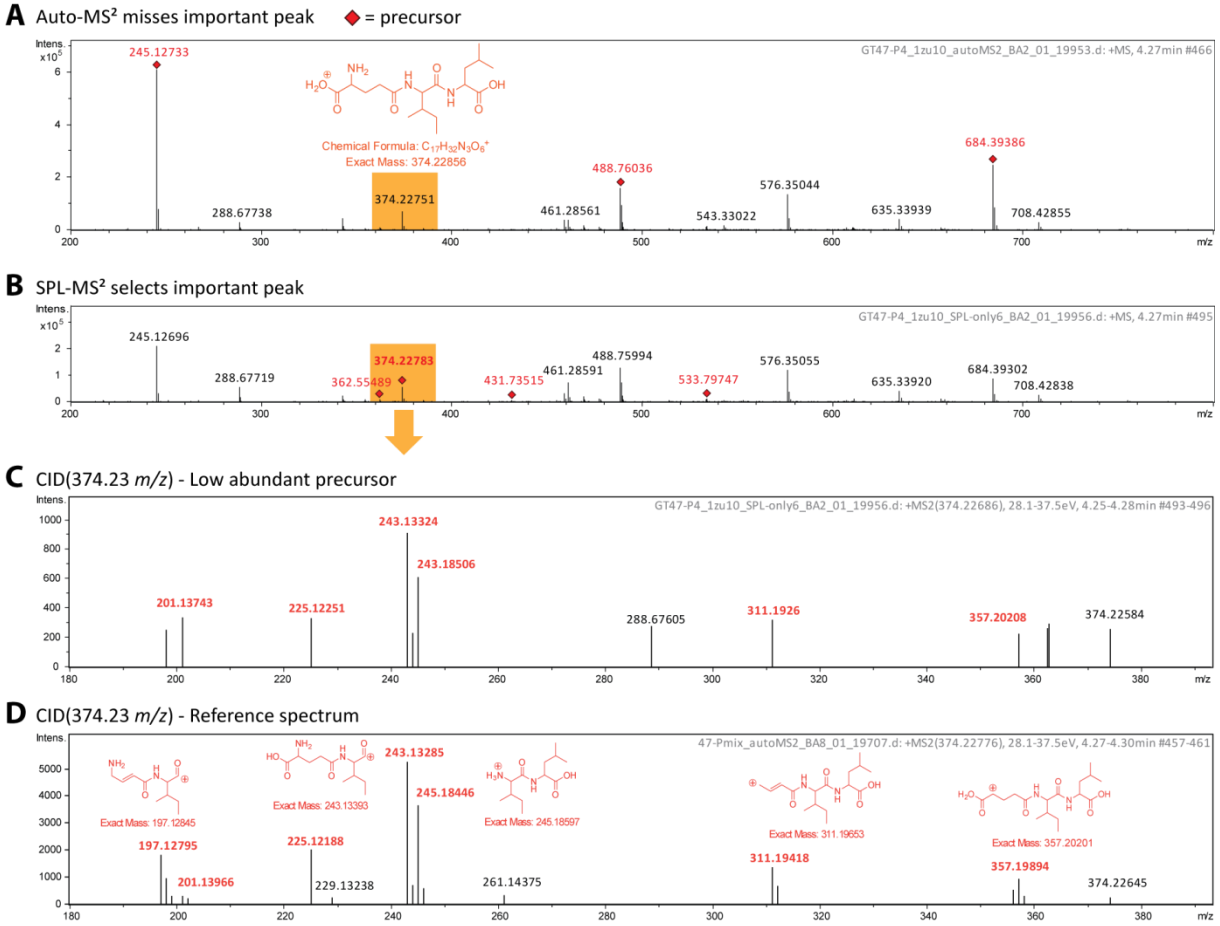


*Figure 23: Analysis of the tripeptide "SH-myxo373". (A) Auto-MS² does not select the low abundant precursor of interest whereas (B) using a SPL-MS² enables MS/MS acquisition. All signals that were selected as precursor for MS/MS analysis are marked red. (C) The MS/MS spectra quality as derived from the low abundant precursor (5.4*10^4 Int.) is evaluated by comparison to a reference spectrum (D). The crude extract of So ceGT47-P was diluted 10-fold with a P-medium blank extract. All m/z values belonging to known fragments are written in red bold letters.*

The spectra quality remains when complex molecules like microsclerodermin M are fragmented (Figure 25). Microsclerodermin M is a PKS/NRPS hybrid compound, which was isolated from So ce38 (Chapter 3). The identity of several fragment ions has been proven by feeding studies and these are marked red in Figure 25. Including several fragments that are derived from water loss, the reference spectrum covers 8 characteristic fragment ions. The fragment spectrum obtained from the SPL-guided acquisition covers 7 out this 8 fragment ions. This is remarkable as the precursor intensity is at the lower limit of what is considered to be reasonable for precursor selection (8900 counts).

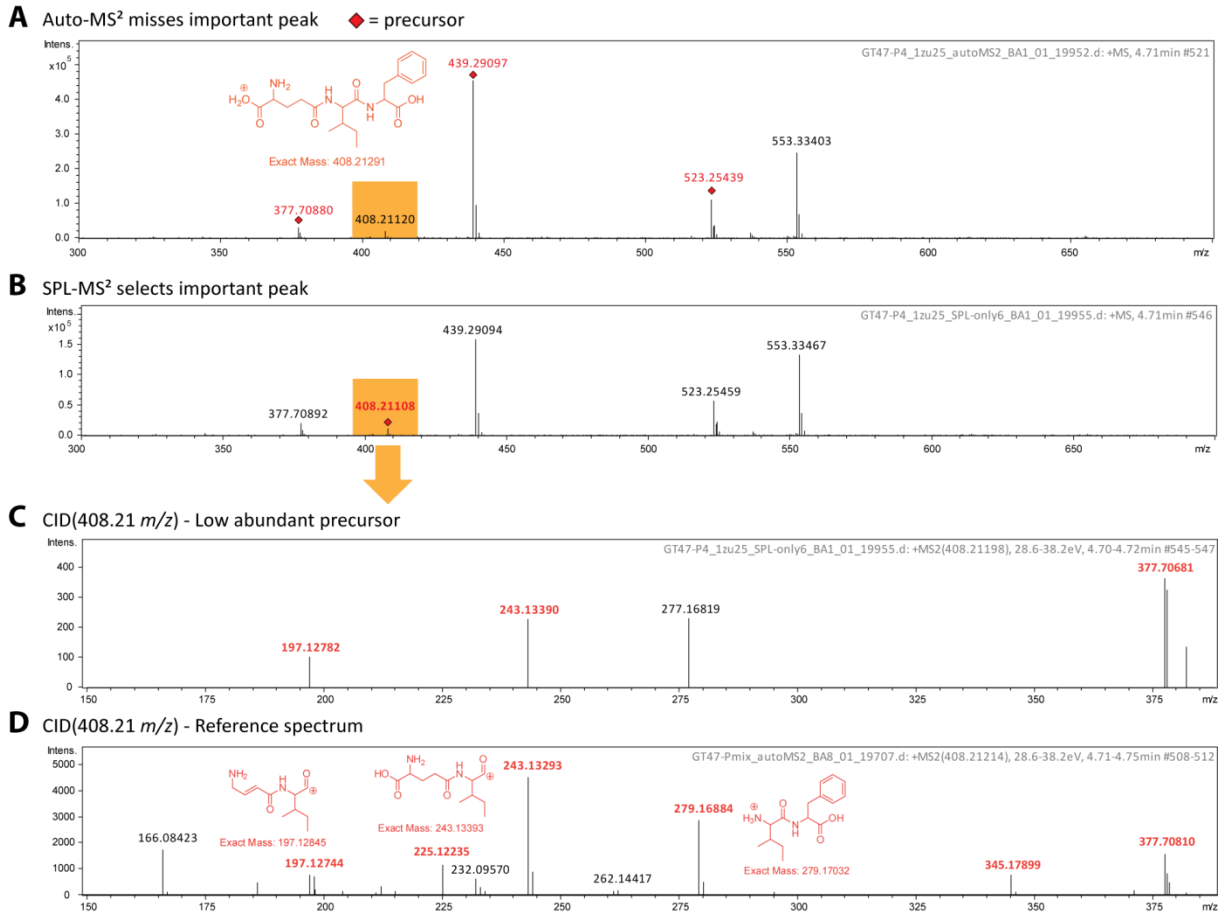


Figure 24: Analysis of the tripeptide "SH-myxo407". (A) Auto-MS² does not select the low abundant precursor of interest whereas (B) using a SPL-MS² enables MS/MS acquisition. All signals that were selected as precursor for MS/MS analysis are marked red. (C) The MS/MS spectra quality as derived from the low abundant precursor ($1.9*10^4$ Int.) is evaluated by comparison to a reference spectrum (D). The crude extract of So ceGT47-P was diluted 25-fold with a P-medium blank extract. All m/z values belonging to known fragments are written in red bold letters.

In light of these results we can conclude that selection of low abundant precursor ions and acquisition of corresponding MS/MS spectra is feasible as reliable fragment spectra can be indeed obtained. In particular, fragment spectra were acquired for three low abundant compounds out of complex crude extracts and each case has proven that the obtained spectra cover a sufficient number of characteristic fragment ions. This result is mandatory when thinking of using this technique of targeted MS/MS-acquisition prior to computational analyses. Computational approaches critically rely on accurate MS and MS/MS data, a request that is feasible using the method developed in this work. Moreover, only MS/MS spectra of "interesting" signals will be used by those downstream processing tools; hence, drawing wrong conclusions should be significantly reduced, e.g. media compounds clustering similar to an interesting compound are out of scope because there will be no MS/MS data available for media compounds. The same is true for algorithms that specifically track peptidic natural

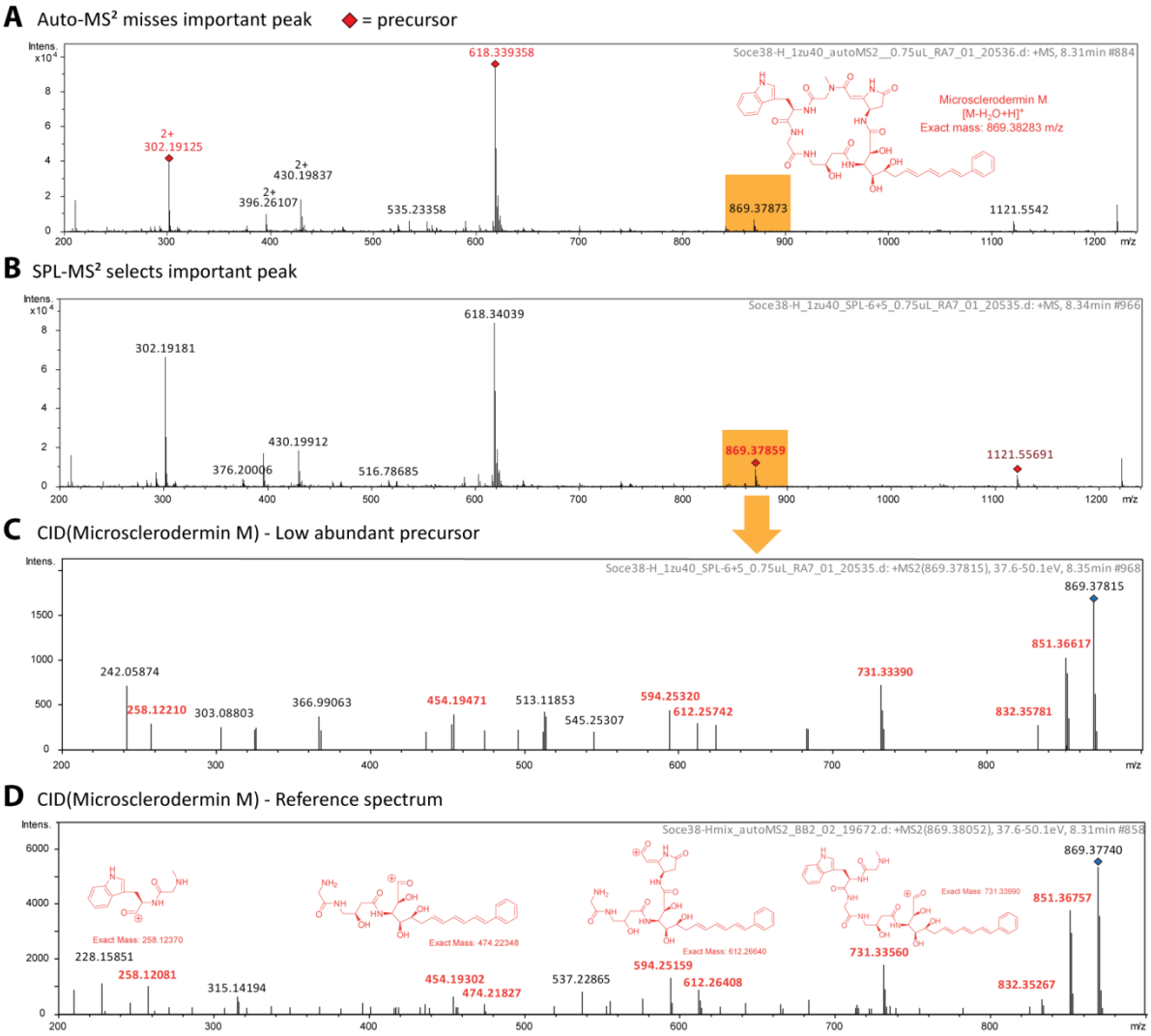


Figure 25: Analysis of microsclerodermin M. (A) Auto-MS² does not select the low abundant precursor of interest whereas (B) using a SPL-MS² enables MS/MS acquisition. All signals that were selected as precursor for MS/MS analysis are marked red. (C) The MS/MS spectra quality as derived from the low abundant precursor (8900 Int.) is evaluated by comparison to a reference spectrum (D). The crude extract of So ce38-H was diluted 53-fold with a H-medium blank extract. All m/z values belonging to known fragments are written in red bold letters.

products. The plethora of nutritional peptides found in complex media is not selected for the acquisition of MS/MS. A detailed evaluation of the method's benefit when using data processing based on the fragmentation tree algorithm[6] is a matter of upcoming research. However, two straight-forward examples of how to benefit from this approach are demonstrated in the two upcoming chapters. First, the SPL is used to identify putatively bioactive compounds in a bioactive fraction of So ceGT47 by highlighting only those compounds that are linked to bacterial metabolism (Chapter 4.5.6). Then, the results of multivariate data analysis supported the fast identification of a new compound class found in So ceGT47 (Chapter 4.5.7).

4.5.6 Revealing Unknown Compounds using the SPL Approach

When crude extracts of *Sorangium cellulosum* So ceGT47 were tested for bioactivity, all of them showed an antimicrobial effect against *Staphylococcus aureus*. Since this activity is of particular interest the crude extract was forwarded to an LC-based fractionation as described in Chapter 4.4 using both, the buffered (pH 6.5) and the acidic eluent system (pH 2.7). The acidic run revealed that ambruticins F and S – both found at similar amounts in all extracts – are active against *S. aureus* which was not known before. In addition another compound with the tentative sum formula $C_{24}H_{33}O_3$ was identified to be active against *S. aureus* as well. This activity was easily linked to a single peak eluting within the designated retention time window. Upon testing fractions of the buffered HPLC run, a single fraction with bioactivity against *S. aureus, M. luteus,* and *M. hiemalis* was identified which was not found in the acidic run before. This could be attributed to an unknown compound which is acid-labile and decomposes upon drying the fractions. How can this particular compound be tracked down with the help of statistics based data evaluation? Is the activity related to an abundant peak or is some low abundant secondary metabolite responsible?

The fact that the activity is found in H and M medium culture helps to identify the compound. The retention time range from the HPLC fractionation is known and the information from statistical data mining can be used to reduce the number of putative compounds that are found within this particular retention time window. For this purpose, the respective SPL from H and M medium experiments are checked for an overlap, i.e. which buckets are found in both SPL. The compound of interest must be part of this "overlap list" as the activity is found in both extracts. The active fraction we are interested in covers the chromatographic region from 7.33 – 7.55 min when using the buffer run (Figure 26 A). Since the statistics data is based on acidic runs, the retention for compounds may be different between both eluent systems. Hence, it is important to allow a larger retention time window in such a case, e.g. a window ranging from 6 to 9 minutes. By comparing both SPL a small subset of only 9 buckets is found in both SPL whereas two of those buckets are relatively high abundant (Figure 26 B). The potentially interesting signals are then located within the buffer run which initially yielded the bioactive fraction in order to check whether those peaks match the bioactive fraction.

Using this methodology, only two signals match the retention time window for which activity was observed: a compound with 241.19156 *m/z* and one with 398.15856 *m/z*. Moreover, MS/MS data for these signals is available as the respective signals are part of the SPL list and hence, selected for the acquisition of fragment spectra. The most abundant signal with 241 *m/z* is a cyclic dipeptide based on accurate mass measurement and fragmentation pattern. These cyclic dipeptides are known to have an antibiotic effect against some Gram positive bacteria including *S. aureus*. The high abundance of this compound is a good indication for its involvement in the observed bioactivity. However, the second

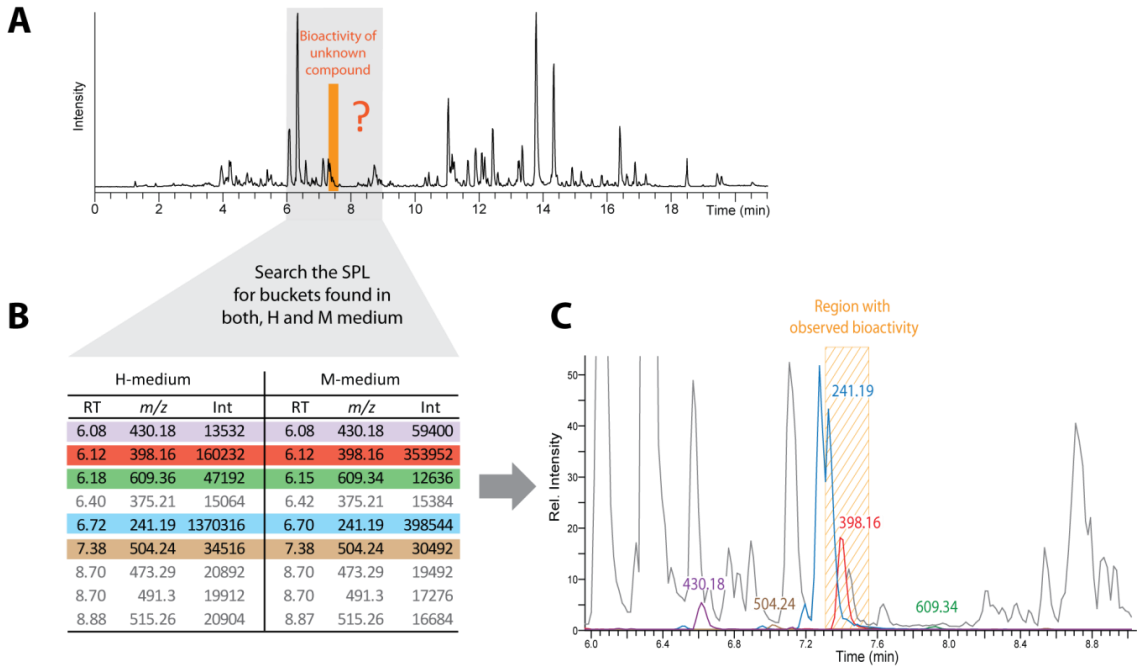


Figure 26: ***A*** *Bioactivity-guided fractionation of H and M-medium cultures of So ceGT47 revealed a fraction being active against* S.aureus*,* M. luteus*, and* M. hiemalis *between 7.33 and 7.55 min in a HPLC run using buffer eluents.* ***B*** *SPL features present in both data sets (H and M medium) were searched within a wide retention time window with the abundant ones highlighted.* ***C*** *The identified m/z values were then searched within the initial fractionation run of* ***A*** *and checked whether they match the active region.*

compound with 398 *m/z* turned out to be a new compound as well. This result would be a good starting point for further approaches toward compound purification and classification. In the end, following this method allows tracking down signals which are most likely responsible for bioactivity. It is moreover no problem to compare buffer and acidic runs even though retention times may vary. The small number of putatively interesting signals makes a search within both data sets achievable and even superior to the manual search without any preliminary information.

4.5.7 Lipothiazoles – A New Compound Class in *S. cellulosum* So ceGT47

A new compound class was identified during this work. The focus on lipothiazoles was initiated by the high impact of lipothiazole-related features in the very first attempts of principal component analysis using extracts of So ceGT47 (Figure 27). This compound family was a good proof of principle for the data mining workflow as explained before. Indeed, the lipothiazole class showed a strong influence in our data mining workflow as depicted in the PCA loadings plot of Figure 27.

In addition, a bioactivity assay of fractionated crude extract indicated a weak activity against *Micrococcus luteus* in this chromatographic region. The two abundant derivatives, lipothiazole A and B were isolated in the course of this work from *S. cellulosum* So ceGT47 extract. The lipothiazoles are NRPS/PKS hybrid structures with a very weak, rather unspecific activity against *Micrococcus luteus* (approx. 100 µg/mL for lipothiazole A). The scaffold of the lipothiazole class features a saturated alkyl chain attached to a thiazole moiety, followed by a short polyketide and two amino acids attached of

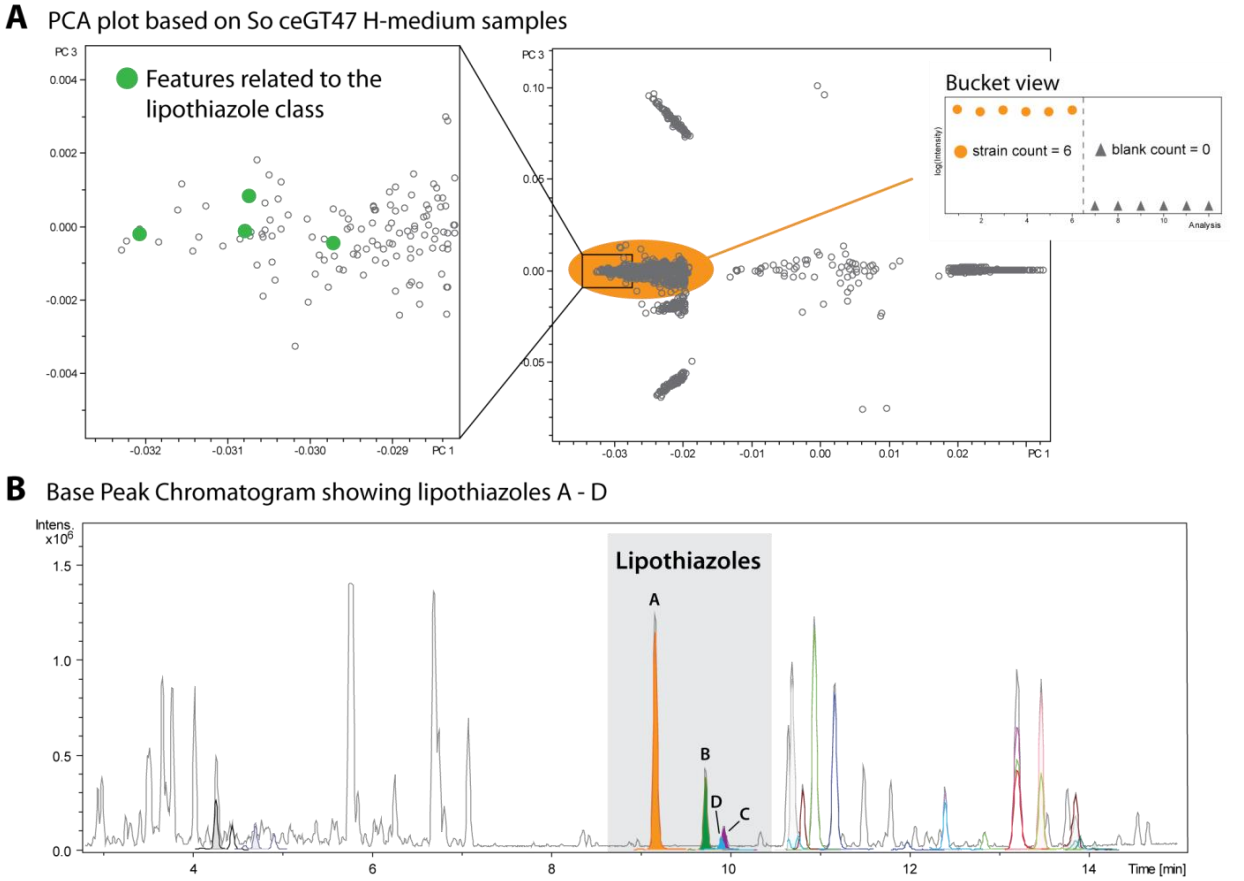


*Figure 27: **A** PCA loadings plot from a So ceGT47 data set. Part of the orange area (strain count = 6 / blank count = 0) are zoomed in order to see the high impact of lipothiazole-related features on the model (green dots). The more left a dot is found, the more impact it has on the overall model. **B** Base peak chromatogram of So ceGT47 – M medium extract highlighting the 4 lipothiazoles*

which serine and alanine are *S*-configured. Lipothiazole A and B were structurally elucidated by NMR, Marfey analysis, and mass spectrometry. HMBC couplings between H2-C3 and H4-C5 confirmed the peptide backbone which is supported by MS/MS fragmentation data (Figure 28). The thiazole moiety was set by HMBC couplings between H9 and C7, C8, and C10 and the chiral carbon of the C-terminal serine is of *S*-configuration as determined by advanced Marfey analysis. Lipothiazole B is almost identical to derivative A and differs solely in the hydroxyl group at the serine. In lipothiazole B, the serine is replaced by an alanine as unambiguously determined by the change in NMR signal shifts of position 18. The identification of lipothiazole B suggests that the serine found in derivative A does originate from an alanine which is hydroxylated in a tailoring step of biosynthesis.

There are additional derivatives identified based on accurate *m/z* measurements and MS/MS analyses. Lipothiazole C seems to be a shunt product of the assembly line. Lipothiazole D has a molecular weight increased by 14 Da compared to lipothiazole A suggesting an additional CH_2 unit somewhere in the molecule. Based on analyses of the different fragmentation patterns we propose the location of this additional CH_2 must be within the alkyl chain. This is reasoned by the fact that fragments

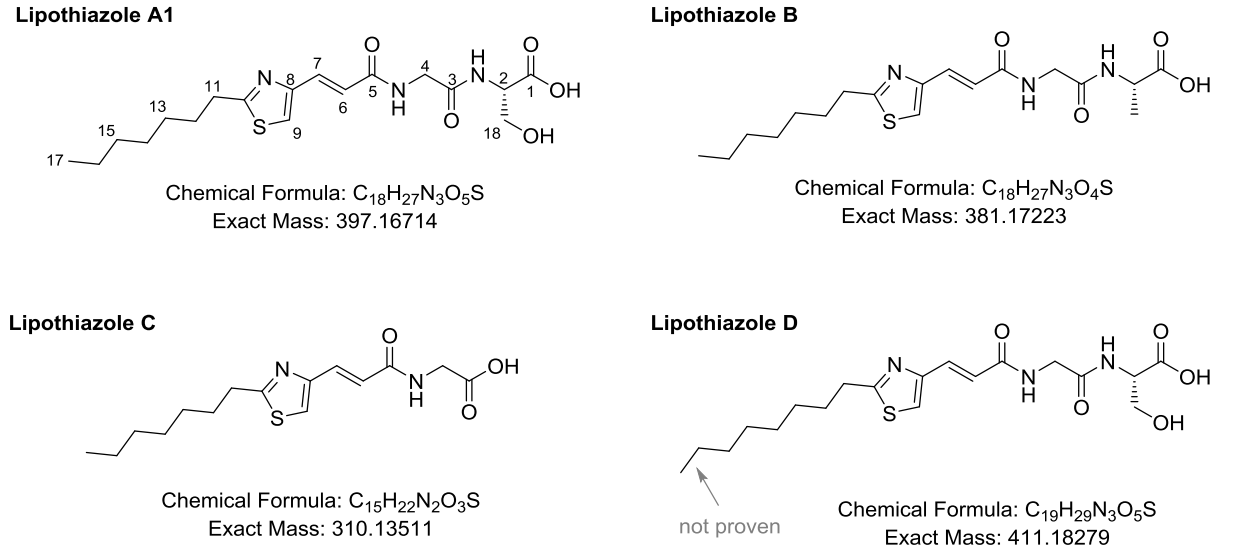


Figure 28: The members of the lipothiazole compound class as found in S. cellulosum So ceGT47. The structure of derivatives C and D is based on the results from MS/MS experiments.

derived from MS^3 analysis are identical for lipothiazole A, B, C, and D (Figure 34). However, the exact position of the additional CH_2 group was not established.

An early evaluation of lipothiazole A fragment spectra suggested the presence of a C-terminal serine followed by a glycine or alternatively the combination alanine followed by glycine (referring to lipothizaole B). With this knowledge we set out to check whether matching NRPS modules are found somewhere in the genome of So ceGT47. For that reason, all biosynthetic gene clusters of So ceGT47 were annotated using antiSMASH2.0. As a consequence of the currently bad sequence quality we were not able to find the cluster at a glance. The information on two NRPS modules was not sufficient. Upon full structure elucidation of lipothiazole It became evident that a thiazole moiety is present in the molecule. These structural elements are usually biosynthesized by special cyclization domains found in NRPS modules that specifically form a 5-membered heterocycle from cysteine whereas the resulting intermediate is frequently oxidized to yield an aromatic heterocycle. With this information in hand, the

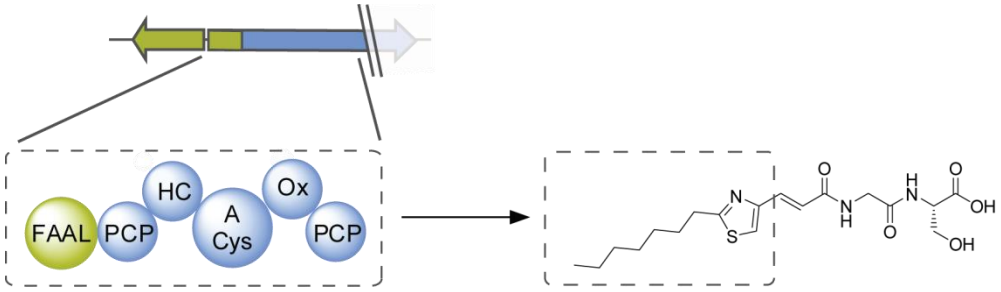


Figure 29:Part of a NRPS-based gene cluster within the S. cellulosum *So ceGT47 genome fits to the eastern part of the lipothiazole scaffold. The cluster fragment was found based on a retrobiosynthetic analysis of lipothiazole following the search for matching NRPS modules. Owing to bad sequence quality, the complete cluster could not be identified.*

genome was searched for an adenylation domain specific for cysteine in combination with a cyclization and an oxidation domain. For this particular query two hits were obtained. The strain So ceGT47 is a known producer of thuggacins which feature a thiazole moiety as well. By comparing the surrounding genes for both query hits one hit could be assigned to thuggacin biosynthesis. Hence, the remaining cluster should be linked to lipothiazole biosynthesis. This notion was supported by the fact that the NRPS module harboring the Cy and HC domain starts with a domain showing homology to fatty acyl adenylate ligases (FAAL). Such a domain could be involved in recruiting the N-terminal alkyl chain found in lipothiazoles (Figure 29). With the currently available genome data the cluster cannot be closed but a full characterization of the biosynthetic machinery behind lipothiazole was not within the scope of this project.

The example of lipothizaole shows how a retrobiosynthetic approach helps to link a compound to the respective gene cluster. However, the ideal situation of using solely the MS/MS data to identify the respective gene cluster like it is proposed by peptidogenomics was not achieved.[22] This puts emphasis on the fact that this approach critically relies on the quality of the available genome data.

4.5.8 Conclusion

Taken together, the established method of statistics-based filtering of LC-MS data is suitable to extract a subset of molecular features that is related to the metabolism of a bacterium. This was done by comparing bacterial strain samples with blank medium samples following the basic assumption that molecular features related to the strain's metabolism are solely found in strain samples. This process of comparing two sample groups was realized by a software-based two-dimensional binning (bucketing) of molecular features with respect to *m/z* and retention time. Features that matched the constraint of only being found in strain samples were extracted from the data set to create a new list. This list covers signals of potential interest as signals originating from secondary metabolites must be related to bacterial metabolism as well. The benefit of such a list becomes obvious when used as the basis for a precursor list for subsequent LC-MS/MS acquisitions. It was clearly shown, that precursor selection is not anymore depending on precursor intensity, a fact of notable importance when thinking of the wide concentration range of metabolites in complex biological samples. The method's efficiency is underpinned by several observations:

1) Measurement of a blank sample under the control of an SPL list results in a low number of MS/MS events related to a basal error level within the SPL and during precursor selection.
2) SPL-MS² of strain extracts diluted with blank extracts maintains a high MS/MS efficiency largely independent of the dilution. This is owing to the fact that high abundant matrix compounds of the blank extract do not affect precursor selection when using a SPL.

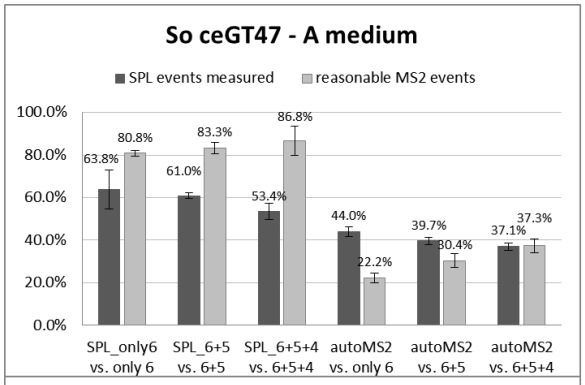
3) The quality of SPL-MS² derived fragment spectra is sufficient even though low abundant precursor ions were used.

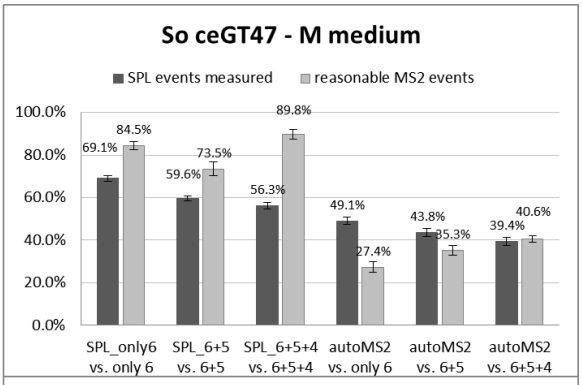
In conclusion, the approach presented in this study has the potential to overcome critical limitations that arise by non-targeted precursor selection in LC-MS/MS acquisition. Modern instrumental platforms such as the hr-QqTOF device used here allow highly accurate and highly resolved MS/MS spectra but acquisition of such data is time consuming. Thus acquisition time should not be wasted for the fragmentation of unneeded precursor ions. This was realized by the approach taken herein. Moreover, the possibility to filter raw data for potentially interesting signals is of interest when searching for new compounds in a bioactivity-based approach. Complex LC-MS data may be a problem when trying to link an observed bioactivity to a certain signal within the LC-MS chromatogram. With respect to fractions covering 10 to 20 seconds of the chromatogram it is laborious to identify *m/z* signals related to bacterial metabolism within such a region. The statistics-based approach supports tracking down peaks of interest as it delivers a selection of *m/z* values of potential interest for every time slot.

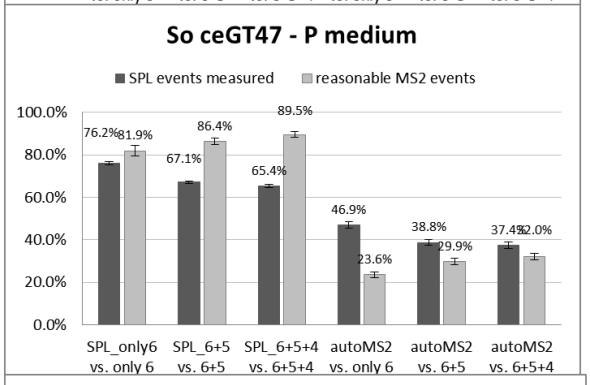
4.6 Supporting Information

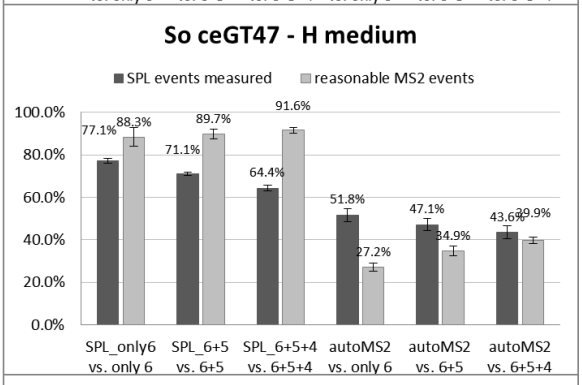
4.6.1 Results of the SPL-based measurements

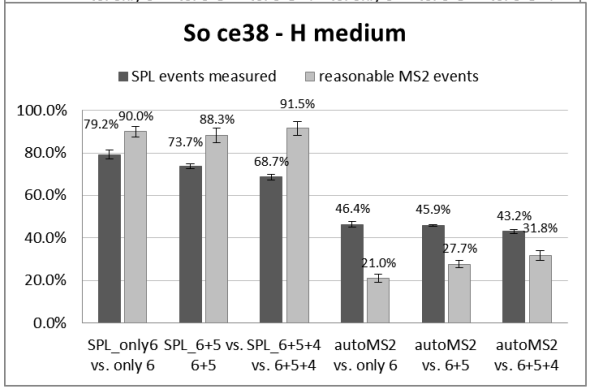
The upcoming graphs show the results of all measurements that were performed under control of SPL lists. The data covers data sets from So ceGT47 and So ce38. For all data sets a similar result is observed which underpins the reproducibility of the observed effects as well as the overall efficacy of the method.

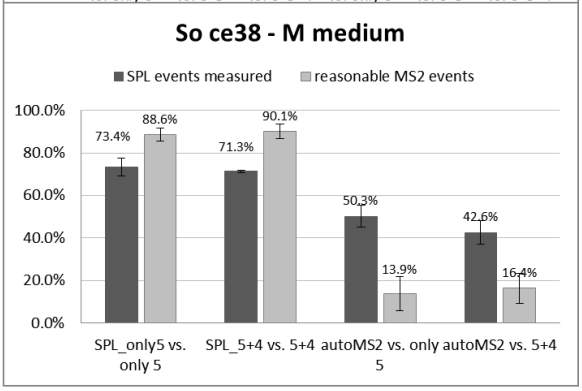


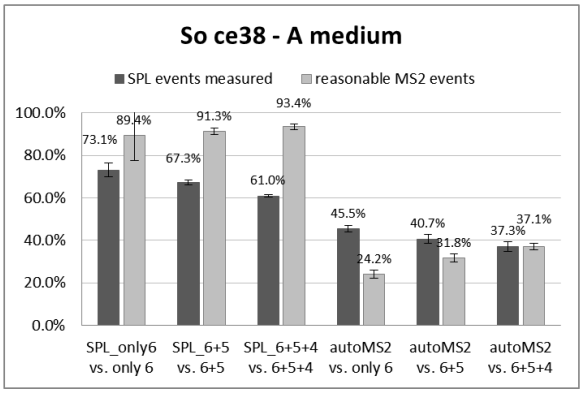












4.6.2 Spiking Results

Table 8 shows the results of the spiking samples measured as triplicate. Values are area values derived from peak integration in the respective extracted ion chromatograms. EICs were created with a 0.1 Da window around the exact *m/z* value. Integration was done automatically but checked manually, though. The purity of the spiked compounds was checked by means of LC-MS and peak integration of all peaks of the base peak chromatogram.

Table 8: Area values (mean value and standard deviation) of the spiked compounds at different concentrations and different crude extracts used. (N=3).

Chlorotonil A	MeOH		So ceGT47-P		So ceGT47-M		So ceGT47-H		So ceGT47-A	
c [µg/mL]	mean	dev	mean	dev	mean	dev	mean	dev	mean	dev
0.04	582	1008	0	0	0	0	0	0	0	0
0.08	4169	509	0	0	0	0	0	0	0	0
0.38	16524	549	0	0	0	0	0	0	0	0
0.76	29585	1242	0	0	0	0	0	0	1893	0
3.80	111306	3314	23657	4439	32462	4736	12291	3661	19909	3545
7.60	205205	35109	69443	4575	86337	9882	45170	10460	78354	9472
38.00	124548	20145	58420	10943	407318	8773	36197	6631	29617	7584

Myxochelin A	MeOH		So ceGT47-P		So ceGT47-M		So ceGT47-H		So ceGT47-A	
c [µg/mL]	mean	dev	mean	dev	mean	dev	mean	dev	mean	dev
0.05	0	0	0	0	0	0	0	0	0	0
0.10	2345	2118	0	0	2029	3514	0	0	0	0
0.50	13868	750	13484	11686	0	3922	0	0	13122	3362
0.99	31493	3762	28388	3717	29286	3922	0	0	27866	0
4.95	131888	4846	126680	15209	129772	14504	77455	4785	103760	20505
9.90	312056	87528	257967	5662	287741	18723	169451	6805	271681	16493
49.50	1363237	36647	1028127	28801	1112899	95758	735288	19818	999281	50773

Sorangicin A	MeOH		So ceGT47-P		So ceGT47-M		So ceGT47-H		So ceGT47-A	
c [µg/mL]	mean	dev	mean	dev	mean	dev	mean	dev	mean	dev
0.05	0	0	0	0	0	0	0	0	0	0
0.10	2834	120	0	0	0	0	2242	2070	0	0
0.48	11877	1677	10138	2848	0	379	8784	6587	13707	1397
0.96	23577	2424	20974	5290	19692	379	19323	1361	20670	0
4.80	97854	7404	98754	2050	93780	3130	86538	4617	76502	4544
9.60	212205	64886	191862	3177	220933	15469	205599	7585	201807	7410
48.00	837681	57944	789297	12809	971202	67613	833516	14151	835399	31137

4.6.3 Lipothiazoles

4.6.3.1 Lipothiazole A

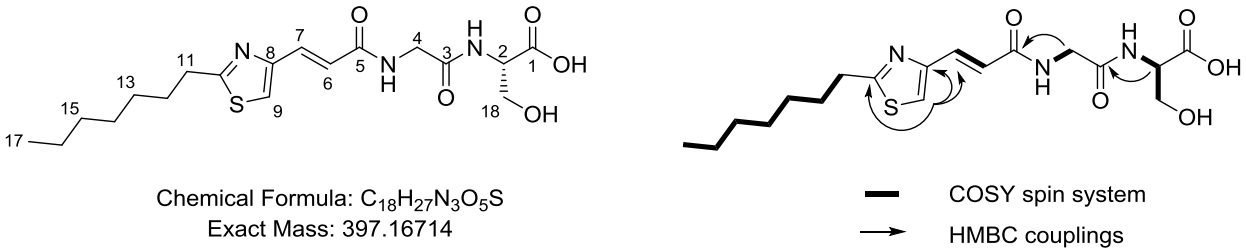
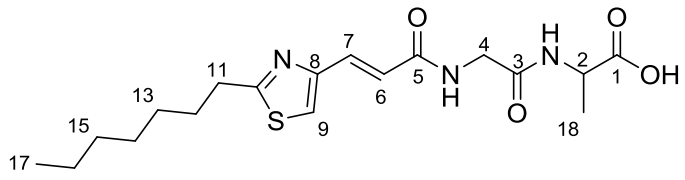


Figure 30: Structure and atom numbering of lipothiazole A (left part) together with HMBC and COSY data mapped onto the structure (right part). The compound was isolated from the myxobacterium S. cellulosum So ceGT47.

	Assignment	δ_C [a]	δ_H (mult., *J* in Hz) [b]	HMBC
Serine	1	173.8	-	-
	2	55.9	4.06 (t, 4.1)	1, 18
	18	62.5	3.85 (dd, 11.3, 3.7) 3.92 (dd, 11,3, 4.6)	1, 2
Glycine	3	171.2	-	-
	4	43.4	4.06 (ov)	3, 5
α,β-unsaturated	5	168.8	-	-
	6	134.1	7.50 (d, 15.3)	5, 7, 8
	7	122.9	6.86 (d, 15.4)	5, 6, 8
Thiazole	8	152.4	-	-
	9	121.8	7.59 (s)	7, 8, 10
	10	173.2	-	-
	11	33.8	3.02 (t, 7.6)	10, 12, 13
	12	30.8	1.81 (m)	11, 13
	13	29.8	ov	ov
	14	29.8	1.39 (m)	ov
	15	32.6	1.31 (m)	ov
	16	23.4	1.32 (m)	ov
	17	14.1	0.91 (t, 6.8)	15, 16

[a] Recorded at 175 MHz, referenced to residual solvent MeOH-d_4 at 49.15 ppm. [b] Recorded at 700 MHz, referenced to residual solvent MeOH-d_4 at 3.31 ppm. [c] Not observed. [ov] overlapping signals.

4.6.3.2 Lipothiazole B



Chemical Formula: $C_{18}H_{27}N_3O_4S$
Exact Mass: 381.17223

Figure 31: Structure and atom numbering of lipothiazole B. The compound was isolated from the myxobacterium S. cellulosum So ceGT47.

	Assignment	δ_C [a]	δ_H (mult., *J* in Hz) [b]	HMBC
Alanine	1	172.0	-	-
	2	49.5	4.41 (d, 7.3)	1, 18
	18	17.9	1.40 (d, 7.3)	1, 2
Glycine	3	173.0	-	-
	4	43.3	4.01 (ov)	3, 5
α,β-unsaturated	5	165.5	-	-
	6	134.1	7.48 (d, 15.4)	5, 7, 8
	7	123.2	6.85 (d, 15.4)	5, 6, 8
Thiazole	8	151.1	-	-
	9	121.5	7.59 (s)	7, 8, 10
	10	171.6	-	-
	11	33.9	3.02 (t, 7.6)	10, 12, 13
	12	30.7	1.81 (m)	11, 13
	13	28.4	ov	ov
	14	28.4	1.39 (m)	ov
	15	32.5	1.31 (m)	ov
	16	23.4	1.40 (m)	ov
	17	14.2	0.91 (t, 6.6)	15, 16

[a] Recorded at 175 MHz, referenced to residual solvent MeOH-d_4 at 49.15 ppm. [b] Recorded at 700 MHz, referenced to residual solvent MeOH-d_4 at 3.31 ppm. [c] Not observed. [ov] overlapping signals.

4.6.3.3 Analytical Data

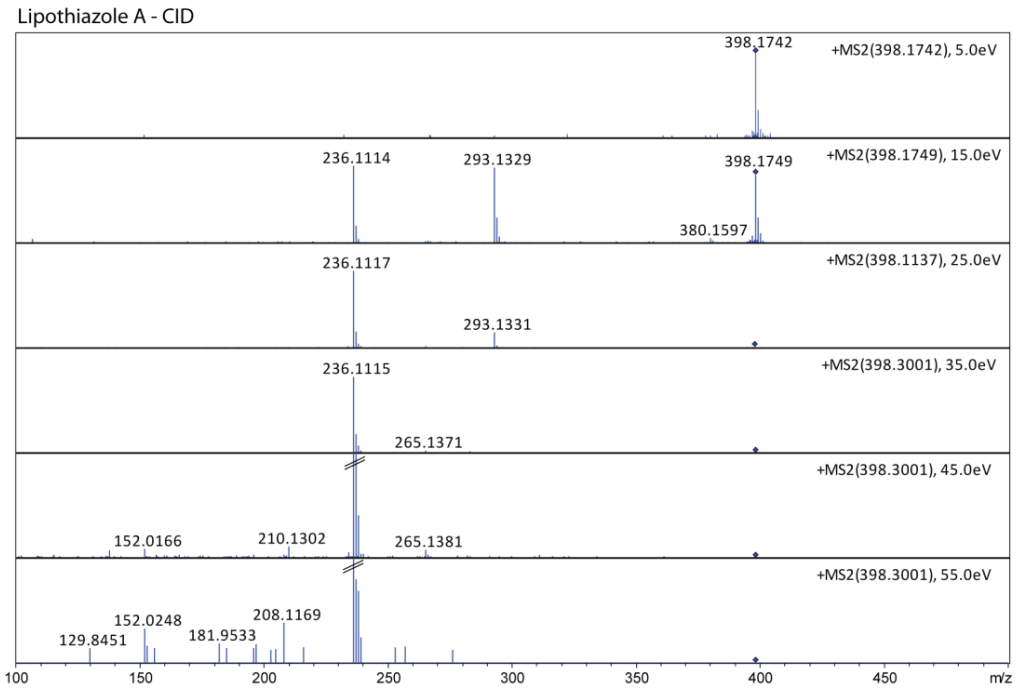


Figure 32:CID spectra of Lipothiazole A using the maXis 4G QqTOF at different fragmentation energies. All spectra are recorded for 1 minute using a direct infusion experiment.

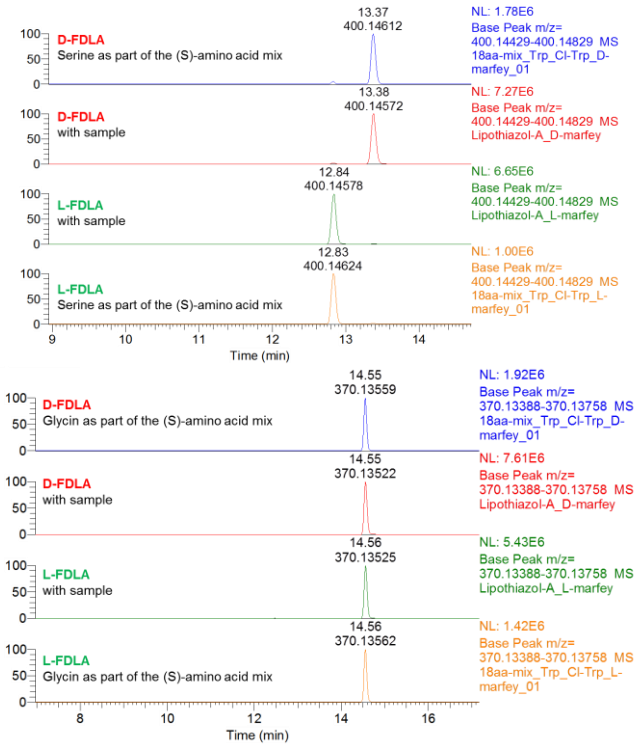


Figure 33:Results of the advanced Marfey analysis of a hydrolyzed sample of lipothiazole A. The serine derivative of Marfey's reagent elutes at the same retention time as the S-configured amino acid reference and thereby proves the presence of a S-serine in lipothiazole A.

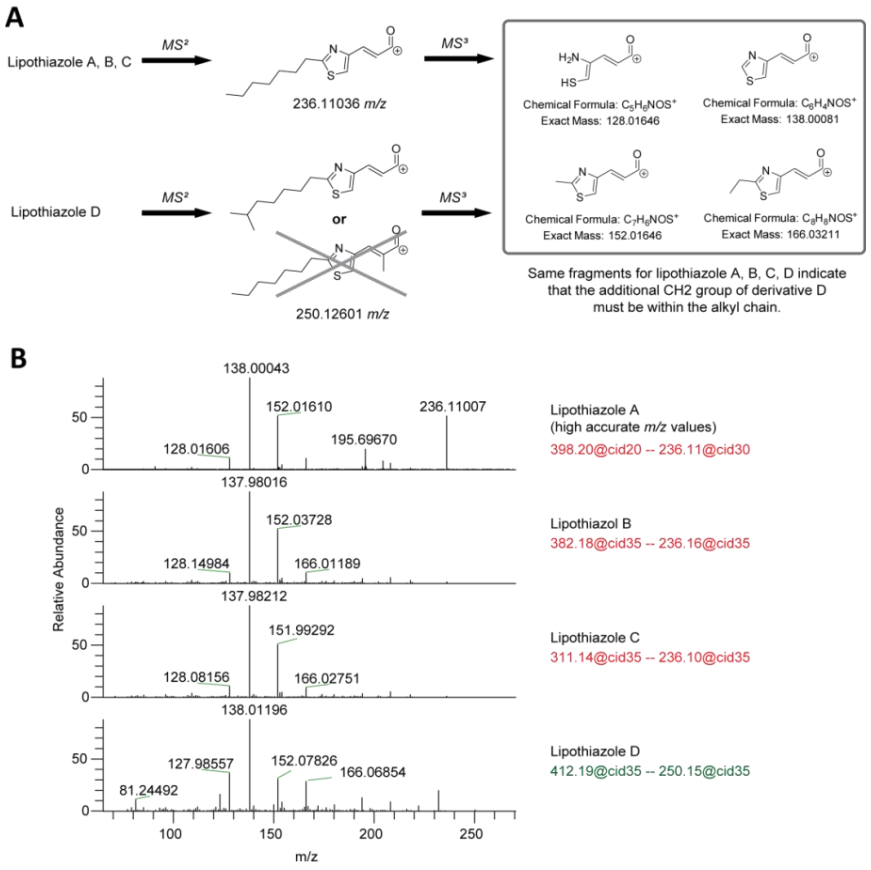


Figure 34: The structure of lipothiazole C and D is concluded from high accurate fragment ions together with the accurate m/z of the precursor ion. All derivatives show a prominent fragment of the western part in MS² (A). Further fragmentation of this western part using MS³ results in identical fragments for all four lipothiazoles (B). This indicates that the additional CH_2 group of lipothiazole D must be located within the alkyl chain.

4.6.3.4 Isolation of Lipothiazoles

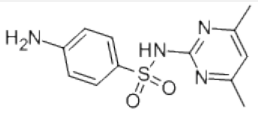
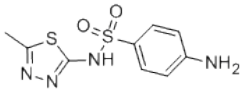
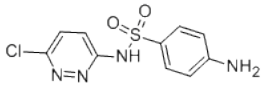
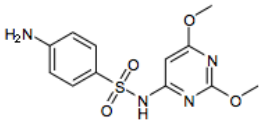
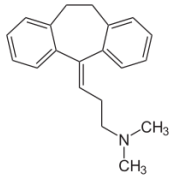
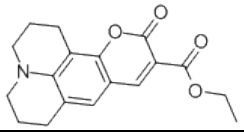
The production medium for So ceGT47 was M medium (10 g/L peptone, 10 g/L maltose, 1 g/L $CaCl_2 \cdot 2H_2O$, 1 g/L $MgSO_4 \cdot 7H_2O$, 8 mg/L Fe-EDTA, 50 mM HEPES, adjusted to pH 7.4 with 10 N KOH). A 5 L shaking flask was used to cultivate 2 L of So ceGT47 using 2 % (w/v) XAD-16 adsorber resin (Rohm & Haas). The culture was harvested after 7 days of fermentation at 30 °C. Cells and XAD was extracted with 4 x 200 mL of methanol. The combined fractions yielded 1.92 g dry weight of crude extract after lyophilization. The crude extract was fractionated using silica flash chromatography with a chloroform/methanol (C/M) eluent system where lipothiazoles elute at 50:50 C/M. The combined fractions were concentrated and subjected to preparative HPLC using a Waters Autopurifier System equipped with a Waters XBridge C18, 150 x 19 mm, 5 µm d_p column operated at room temperature. The eluent was (A) H_2O + 0.1 % FA and (B) ACN + 0.1 % FA delivered at 25.5 mL/min. The gradient started at 43 % B, increased to 54 % B in 6 min and to 95 % B in another 0.5 min where it remained for 1.5 min for column flushing. Since there was an unknown plasticizer in our fractions, another purification step was necessary. The combined fractions of interest were dried, dissolved in methanol, and forwarded to a semipreparative Dionex HPLC system (P680 pump, TCC100 thermostat, and PDA100 detector) equipped

with a Phenomenex Luna C18, 250 x 4.6 mm, 4 µm d_p column. Separation was achieved by a linear gradient using (A) H_2O + 0.1 % FA and (B) ACN+ 0.1 % FA at a flow rate of 5 mL/min and 30 °C. The gradient started at 45 % B and increased to 80 % B in 15 min where it remained for 1 min before reequilibration under initial conditions. UV data were acquired at 280 nm. A maximum of 100 µL of the sample was manually injected before fraction collection, yielding 1.2 mg of lipothiazole A and 0.7 mg of lipothiazole B as white amorphous solid.

4.6.1 LC-MS Test Mix

The test mix is injected at least after every 15 samples (V_{inj} = 1 µl). In well plate sequences the test mix is measured after each row. The data is compared to a reference which was acquired with a new column. Deviations in retention time must be less than 0.1 minutes with a perfect peak shape.

Table 9:Components of the LC-MS test mix at HIPS-MINS, Saarland University. The test mix is dissolved in H_2O/MeOH mixture of 9:1 (v/v).

Compound	Structure	Sum formula	MW [g/mol]	CAS	Conc. [mg/L]
Sulfamethazine		C12H14N4O2S	278.33	57-68-1	10
Sulfamethizole		C9H10N4O2S2	270.33	144-82-1	10
Sulfachloropyridazine		C10H9ClN4O2S	284.72	80-32-0	10
Sulfadimethoxine		C12H14N4O4S	310.33	122-11-2	10
Amitriptyline		C20H23N*HCl	313.9	549-18-8	10
Coumarin-314		C18H19NO4	313.4	55804-66-5	20

Amitriptyline is important to track the behavior of basic compounds in our standard screening. This is important to monitor as it may change although other peaks remain constant. Coumarin-314 is strong fluorescent dye which allows tracking of the fractionated peak. Even very low amounts of coumarin-314 can be detected by UV irradiation. All components are detected at 220 nm and (+)-ESI-MS. A stock solution of the sulfa drugs (c = 2000 µg/mL) is prepared in water/methanol 1:1 (v/v). Coumarin-314 is diluted in methanol to a concentration of 1000 µg/mL. Amitriptyline is dissolved in water to a concentration of 1000 µg/mL. A working solution is prepared in a 10 ml volumetric flask, which is filled

with 8 mL of water and 1 mL of methanol. To this is added: 50 µL of sulfa drugs stock, 100 µL of amitriptyline stock, and 200 µL of coumarin-314 stock. The flask is filled with water to end up with the final concentrations as given in Table 9. The final solution is in a water/methanol mixture in ratio 9:1 (v/v). Higher amounts of methanol must be avoided as this will alter retention of the early eluting sulfa drugs. The mix is stable for months when stored in glass vials in a dark place (autosampler rack).

4.7 Experimental Section

4.7.1 Culture Conditions and Extraction

Experiments are based on two *Sorangium cellulosum* strains, So ceGT47 and So ce38. Fresh cultures were received from cryogenic stocks and used to inoculate M medium cultures. These pre-cultures inoculated new cultures but using four different media (M, P, H, A). Cultures were transferred once again to fresh media of the same type to guarantee a full adaption of each strain to the nutrition conditions. Such adapted cultures are then used to prepare XAD cultures for screening. XAD is a PS-DVB-based adsorber resin (XAD-16, Rohm & Haas) that is supplemented to the liquid broth in order to adsorb secondary metabolites. Cultivation was done in 300 mL shaking flasks filled with 45 mL of medium, 1 mL of XAD-16 slurry (1 % w/v, Rohm & Haas) and 4 mL of inoculation culture. In case of “blank” media cultivations, 49 mL of medium and 1 mL of XAD-16 slurry were used. The flasks were incubated at 30 °C and 140 rpm for 6 days (So ceGT47) or 8 days (So ce38), respectively. After the incubation time cultures were harvested, centrifuged and the adsorber resin together with cells extracted using 2 x 25 mL of methanol. The combined organic extracts were filtered and dried using a rotational evaporator. Residues were suspended in 1 mL methanol, which corresponds to a 50-fold concentration compared to cultivation. Samples need to be centrifuged to obtain a clear supernatant, which is used for LC-MS. Each strain/medium combination was prepared as three independent biological replicates.

4.7.2 Cultivation Media

A medium. Glycerin 4 g/L, soluble starch 8 g/L, soy flour 4 g/L, yeast extract 2 g/L, $CaCl_2 \cdot 2H_2O$ 1 g/L, $MgSO_4 \cdot 7H_2O$ 1 g/L, 50 mM HEPES 11.9 g/L, Fe-EDTA 8 mg/L. The pH value is adjusted to 7.4 using 10 M KOH.

H medium. Glucose 2 g/L, soluble starch 8 g/L, soy flour 2 g/L, yeast extract 2 g/L, $CaCl_2 \cdot 2H_2O$ 1 g/L, $MgSO_4 \cdot 7H_2O$ 1 g/L, 50 mM HEPES 11.9 g/L, Fe-EDTA 8 mg/L. The pH value is adjusted to 7.4 using 10 M KOH.

M medium. Maltose 10 g/L, soy peptone 10 g/L, $CaCl_2 \cdot 2H_2O$ 1 g/L, $MgSO_4 \cdot 7H_2O$ 1 g/L, 50 mM HEPES 11.9 g/L, Fe-EDTA 8 mg/L. The pH value is adjusted to 7.4 using 10 M KOH.

P medium. Peptone 2 g/L, soluble starch 8 g/L, probion 4 g/L, yeast extract 2 g/L, $CaCl_2 \cdot 2H_2O$ 1 g/L, $MgSO_4 \cdot 7H_2O$ 1 g/L, 50 mM HEPES 11.9 g/L, Fe-EDTA 8 mg/L. The pH value is adjusted to 7.5 using 10 M KOH.

4.7.3 Pre-cleaning of Adsorber Resin XAD-16

The adsorber resin product *Amberlite XAD16* (Rohm & Haas) is shipped under slightly wet conditions of pH 10 with high amounts of NaCl and Na_2CO_3. In addition, there are polymeric impurities, which affect the LC-MS screening. In order to clean the adsorber resin it is necessary to adjust the pH value to neutral conditions and wash off the salts. Furthermore, polymeric impurities are removed by two consecutive Soxhlet extractions with acetone and methanol. Any solvent residues are removed by thoroughly washing with water prior to autoclaving aqueous slurries of XAD-16 to yield a very clean product suitable for screening purposes.

4.7.4 Standardized LC-MS Screening Method

All measurements 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 d_p column (Waters, Eschborn, Germany). One µl sample is separated 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 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 width. The LC flow is split to 75 µL/min 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; dry gas temperature, 200 °C. Ion transfer parameters were: funnelRF, 350 Vpp; multipoleRF, 400 Vpp; quadrupole ion energy, 5 eV @ low *m/z* 200. Collision cell is set to 8 eV with a collisionRF of 2500 Vpp in full scan mode. Ion cooler settings are: transfer time, 90 µs; ion cooler RF, 120 Vpp; pre puls storage, 5 µs. 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, a lock

mass (Agilent Chip Cube High Mass HP-1221, Art.# G1982-85001) is present in the ion source all the time and allows recalibration of each spectrum.

4.7.5 Different settings for MS/MS acquisition

The MS/MS methods used in this study are identical to the standard LC-MS screening method in terms of chromatography, ion source and ion transfer parameters. They differ only in some MS/MS parameters (see Table 10) and in the ion cooler settings.

Table 10: Different MS/MS methods as used in this study. Changes are related to acquisition speed, number of precursors and the way precursors are selected, i.e. automatically based on most abundant ones (auto-MS²) or based on a precursor list (SPL, scheduled precursor list).

#	mode		precursor	acq.speed #1 [Hz]	acq.speed #2 [Hz]	acq.int 1	acq.int 2	duty cycle [s]
1	auto	MS²	2	1.5	2.9	1E+04	5E+04	1.2 - 1.8
2	auto	MS²	3	1.5	2.9	1E+04	5E+04	1.5 - 2.4
3	auto	MS²_fast	3	1.5	6.7	1E+04	1E+05	0.9 - 2.4
4	SPL	2prec	2	1.5	2.9	1E+04	5E+04	1.2 - 1.8
5	SPL	3prec	3	1.5	2.9	1E+04	5E+04	1.5 - 2.4
6	SPL	3prec_fast	3	2.0	6.7	1E+04	1E+05	0.9 - 2.0
7	SPL	4prec_fast	4	2.0	6.7	1E+04	1E+05	1.1 - 2.5
8	SPL	strict	all	2.0	10.0	1E+04	1E+05	-

Ion cooler settings for MS/MS mode do timely vary within a single MS/MS acquisition (ion cooler sweeping). Ion cooler RF sweeps from 120 to 80 Vpp while CID energy sweeps from 90 to 120 % of the set value. Both sweeps are done in 50/50 timing (Table 11).

Table 11: Ion cooler sweeping during a MS/MS scan. Ion cooler RF reduction allows better detection of low m/z fragments.

Time [%]	IC RF [Vpp]	Coll. Energy [%]
0	120	90
25	120	120
50	80	120
75	80	90

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 minutes.

4.7.6 Method for Fractionation of Crude Extracts

4.7.6.1 Fractionation using Acidic Eluent

Fractionation was performed on a Dionex Ultimate 3000 RSLC system using a Waters BEH C18, 100 x 2.1 mm, 1.7 µm d_p column by injection of 5 µl methanolic sample. Separation was achieved by a linear gradient with (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 550 µL/min and 45 °C. The gradient was initiated by a 0.27 min isocratic step at 5 % B, followed by an increase to 95 % B in 18 min to end up with a 1.5 min flush step at 95 % B before reequilibration under the initial conditions. Coupling the HPLC to the MS was supported by an Advion Triversa Nanomate nano-ESI system attached to an Orbitrap (Thermo Scientific, Dreieich, Germany). Mass spectra were acquired in centroid mode ranging from 200 – 2000 *m/z* at a resolution of R = 30000. The flow was split to 500 nL/min before entering the ion source. The Nanomate transfers the remaining LC flow into a 96 well plate within the time range from 0.6 to 20.6 min. Each well is filled for approx. 0.22 min (120 µl) yielding 92 wells in summary. The well plate is dried using a Genevac centrifugal evaporator (Genevac Ltd., Suffolk, UK) for at least 3 h at 30 °C and reduced pressure. The dry plate is used for bioactivity assay afterwards.

4.7.6.2 Fractionation using Buffered Eluent

For buffered fractionation the same setup as described in 4.7.6.1 is used. The eluents are changed to (A) 5 mM NH_4HCOO in H_2O and (B) 5 mM NH_4HCOO in ACN/H_2O 9:1 (v:v). The pH is not adjusted. Gradient separation is initiated with a 0.27 min isocratic step at 5 % B, followed by an increase to 100 % B in 18 min to end up with a 1.5 min step at 100 % B before reequilibration under the initial conditions.

4.7.7 Marfey Analysis

4.7.7.1 Marfey derivatization protocol

Marfey derivatization was performed with NMR, ozonolysis and diol-cleavage samples of the compounds. The protocol is as follows:

- Put 50 µg of sample into a 1.4 mL glass vial
- Add 100 µL of 6 N HCl. Fill the vial with nitrogen and close it. Keep it at 110 °C for 45 minutes. Open the vial to let the contents dry at 110 °C for another 15 minutes. Do not exceed 60 min as tryptophan easily decomposes (if present).
- Dissolve the residues in 110 µL H_2O. Prepare two 1.5 mL PP tubes and add 50 µL of the aqueous solution in each.
- Add 20 µL of 1 N $NaHCO_3$ in each tube (which will adjust the pH to approx. 9)
- Add 20 µL of 1 % (w/v) Marfey's reagent in acetone: D-FDLA or L-FDLA, respectively.
- Keep for 1 hour at 40 °C, 700 rpm.
- Add 10 µL of 2 N HCl to stop reaction and 300 µL of ACN to end up with 400 µL total volume.
- Centrifuge the sample and measure the supernatant by HPLC-MS

4.7.7.2 HPLC-MS analysis of Marfey samples

All measurements were performed on a Dionex Ultimate 3000 RSLC system using a Waters BEH C18, 100 x 2.1 mm, 1.7 µm d_p column by injection of 1 µl sample. Separation was achieved by a gradient using (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 550 µl/min and 45 °C. The gradient was as follows: starting at 5 % B to increase to 10 % B in 1 min, from 1 to 15 min increase to 35 % B, from 15 to 22 min increase to 50 % B, from 22 to 25 min increase to 80 % B. After a 1 min hold at 80 % B the system was reequilibrated with initial conditions for 5 minutes. UV data was acquired at 340 ± 8 nm and MS detection was performed simultaneously. Coupling the HPLC to the MS was supported by an Advion Triversa Nanomate nano-ESI system attached to a Thermo Fisher Orbitrap. LC flow is split to 500 nL/min before entering the ion source. Mass spectra were acquired in centroid mode ranging from 150 – 1000 m/z at a resolution of R = 30000.

4.7.8 Bioactivity assay

4.7.8.1 Bacterial Cultures

All microorganisms were handled under standard conditions recommended by the depositor. Overnight cultures of microorganisms were prepared in EBS medium (0.5% peptone casein, 0.5% protease peptone, 0.1% peptone meat, 0.1% yeast extract; pH 7.0) or TSB medium (1.7% peptone casein, 0.3% peptone soymeal, 0.25% glucose, 0.5% NaCl, 0.25% K_2HPO_4; pH 7.3). The latter medium was used for *E. faecalis* and *S. pneumonia* cultures. Yeast and fungi were grown in Myc medium (1% phytone peptone, 1% glucose, 50 mM HEPES, pH 7.0).

4.7.8.2 Microbial Susceptibility Assay (MIC)

Overnight cultures of microorganisms were diluted to OD_{600} 0.01 (bacteria) or 0.05 (yeast/fungi) in the respective growth medium. Serial dilutions of compounds were prepared as duplicates in sterile 96-well plates. The cell suspension was added and microorganisms were grown overnight on a microplate shaker (750 rpm, 30°C or 37°C). Growth inhibition was assessed by measuring the OD_{600} on a plate reader and supported by manual evaluation. MIC_{50} values were determined as average relative to respective control samples by sigmoidal curve fitting.

4.7.9 Spiking of Myxobacterial Crude Extracts

Methanolic crude extracts of *S. cellulosum* So ceGT47 were spiked with a 3-compound mixture containing myxochelin A, sorangicin A, and chlorotonil A. Stock solutions (2500 µg/mL) of myxochelin A and sorangicin A were prepared in methanol while chlorotonil A was dissolved in chloroform. A 100 µL of each stock were mixed with 200 µL of methanol to give a mixture that contains 500 µg/mL of each

compound. This main stock solution (#1, Table 12) is subsequently diluted with methanol to give another 6 solutions.

Table 12: Dilution scheme as used for the preparation of spiking samples.

c_stock [µg/mL]	3-cpd-mix	MeOH [µl]	final volume [µl]
500	-	-	-
100	#1, 100 µl	400	500
50	#2, 100 µl	100	200
10	#2, 100 µl	900	1000
5	#3, 200 µl	200	400
1	#4, 100 µl	400	500
0.5	#5, 200 µl	200	400

Five µL of each stock (#1-7, Table 12) is mixed with 45 µL crude extract. All pipetting was performed in a temperature-controlled room maintained at 4 °C in order to avoid evaporation of the highly volatile solvents. LC-MS analysis of 1 mg/mL samples were used to determine the compound purities of myxochelin A (99 %), sorangicin A (96 %), and chlorotonil A (76 %). All samples were measured as triplicate using the standard LC-MS screening method with an injection volume of 1 µL.

4.8 References

(1) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2012**, *75*, 311–35.
(2) Boucher, H. W.; Talbot, G. H.; Bradley, J. S.; Edwards, J. E.; Gilbert, D.; Rice, L. B.; Scheld, M.; Spellberg, B.; Bartlett, J. *Clin. Infect. Dis.* **2009**, *48*, 1–12.
(3) Lewis, K. *Nat. Rev. Drug Discov.* **2013**, *12*, 371–87.
(4) Bérdy, J. *J. Antibiot. (Tokyo).* **2005**, *58*, 1–26.
(5) Scheubert, K.; Hufsky, F.; Böcker, S. *J. Cheminform.* **2013**, *5*, 12.
(6) Rasche, F.; Scheubert, K.; Hufsky, F.; Zichner, T.; Kai, M.; Svatoš, A.; Böcker, S. *Anal. Chem.* **2012**, *84*, 3417–26.
(7) Rojas-Cherto, M.; Peironcely, J. E.; Kasper, P. T.; van der Hooft, J. J. J.; de Vos, R. C. H.; Vreeken, R.; Hankemeier, T.; Reijmers, T. *Anal. Chem.* **2012**, *84*, 5524–34.
(8) Ng, J.; Bandeira, N.; Liu, W.-T.; Ghassemian, M.; Simmons, T. L.; Gerwick, W. H.; Linington, R.; Dorrestein, P. C.; Pevzner, P. A. *Nat. Methods* **2009**, *6*, 596–9.
(9) Kavan, D.; Kuzma, M.; Lemr, K.; Schug, K. A.; Havlicek, V. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1177–1184.
(10) Guthals, A.; Watrous, J. D.; Dorrestein, P. C.; Bandeira, N. *Mol. Biosyst.* **2012**, *8*, 2535–44.
(11) Nguyen, D. D.; Wu, C.-H.; Moree, W. J.; Lamsa, A.; Medema, M. H.; Zhao, X.; Gavilan, R. G.; Aparicio, M.; Atencio, L.; Jackson, C.; Ballesteros, J.; Sanchez, J.; Watrous, J. D.; Phelan, V. V; van de Wiel, C.; Kersten, R. D.; Mehnaz, S.; De Mot, R.; Shank, E. a; Charusanti, P.; Nagarajan, H.; Duggan, B. M.; Moore, B. S.; Bandeira, N.; Palsson, B. Ø.; Pogliano, K.; Gutiérrez, M.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E2611–20.
(12) Cortina, N. S.; Krug, D.; Plaza, A.; Revermann, O.; Müller, R. *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 811–6.
(13) Hou, Y.; Braun, D. R.; Michel, C. R.; Klassen, J. L.; Adnani, N.; Wyche, T. P.; Bugni, T. S. *Anal. Chem.* **2012**, *84*, 4277–83.
(14) Farag, M. A.; Weigend, M.; Luebert, F.; Brokamp, G.; Wessjohann, L. A. *Phytochemistry* **2013**.
(15) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2006**, *7*, 234.
(16) Höfle, G.; Steinmetz, H.; Gerth, K.; Reichenbach, H. *Liebigs Ann. der Chemie* **1991**, *1991*, 941–945.
(17) Irschik, H.; Reichenbach, H.; Höfle, G.; Jansen, R. *J. Antibiot. (Tokyo).* **2007**, *60*, 733–8.
(18) Buntin, K.; Irschik, H.; Weissman, K. J.; Luxenburger, E.; Blöcker, H.; Müller, R. *Chem. Biol.* **2010**, *17*, 342–56.
(19) Rasche, F.; Scheubert, K.; Hufsky, F.; Zichner, T.; Kai, M.; Svatoš, A.; Böcker, S. *Anal. Chem.* **2012**, *84*, 3417–26.
(20) Van den Berg, R. a; Hoefsloot, H. C. J.; Westerhuis, J. a; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142.
(21) Eriksson, L.; Antti, H.; Gottfries, J.; Holmes, E.; Johansson, E.; Lindgren, F.; Long, I.; Lundstedt, T.; Trygg, J.; Wold, S. *Anal. Bioanal. Chem.* **2004**, *380*, 419–29.
(22) Kersten, R. D.; Yang, Y.-L.; Xu, Y.; Cimermancic, P.; Nam, S.-J.; Fenical, W.; Fischbach, M. A.; Moore, B. S.; Dorrestein, P. C. *Nat. Chem. Biol.* **2011**, *7*, 794–802.

5 Discussion

This thesis covers several aspects of natural product research. On the one hand it deals with isolation and structure elucidation of three secondary metabolite classes from *Sorangium cellulosum* So ce38 and So ceGT47 alongside with the characterization of the underlying biosynthetic machinery. On the other hand, an analytical approach is presented which addresses the high complexity of LC-MS data that is frequently observed for myxobacterial crude extracts. In detail, a new approach for precursor selection in LC-MS/MS analyses is achieved which increases the informational content of the acquired data. Both aspects can be advantageously combined as they contribute to complementary approaches of natural product research that emerged with the increasing availability of whole genome data over the last decade. These data lead to a tremendous increase in characterized biosynthetic gene clusters, which in turn initiated the development of computational methods to characterize compounds on gene level. In principle, two different methods of how a compound can be linked to its gene cluster exist:

- The compound-to-gene approach is based on a known or partially known compound and uses retrobiosynthetic considerations and compound-specific structural elements (e.g. amino acid building blocks, halogenations, *et cetera*) to link a given compound to a matching biosynthetic cluster.

- The gene-to-compound approach starts with *in silico* analysis of genomes aiming toward detection of biosynthetic gene clusters which can facilitate the prediction of a molecular structure and/or the attempt to create a knock-out mutant of a given cluster in order to link the gene cluster to a compound.

The overall idea aims toward connecting both approaches in order to link a gene cluster to its compound and vice versa, thereby supporting the fast characterization of new compounds and their underlying biosynthesis. Such a project is rationalized by the feasibility of the compound-to-gene approach at the stage of LC-MS analysis. In particular, the combination of accurate *m/z* measurements and fragmentation data of secondary metabolites holds information about structural features, e.g. amino acid building blocks. With state-of-the-art mass spectrometric devices, these data are reliably accessible from crude extracts when analyzed by LC-MS/MS. This information can facilitate either the identification of a compound or help to classify it as member of a certain compound class. Similarly, the gene-to-compound approach can provide structural information based on the biosynthetic modules

identified in a given genome. At this point the approaches connect as both determine structural features derived from different types of analysis.

Such an approach depends on two crucial aspects, in particular the reliability of computational models for the *in silico* characterization of biosynthetic gene cluster *and* the informational content of LC-MS data. This thesis supports this project by contributing to both aspects: the identification and in depth characterization of two gene clusters helps to improve existing computational models, whereas the introduction of an improved LC-MS/MS workflow increases the informational content of acquired analytical data.

5.1 Predicting Secondary Metabolite Structures – Discrepancies between Genome-based *in silico* Analysis and Real Structure

With the feasibility of whole genome sequencing, an ever-growing importance is given to *in silico* based analysis of genomes. The past years have seen tremendous advances in using genome data to characterize biosynthetic pathways but some methods related to the characterization of biosynthetic gene cluster still suffer from weaknesses as discussed in the following. While the initial identification of genomic loci harboring NRPS, PKS, or other biosynthetic genes is based on powerful tools of high reliability, subsequent steps of specifically addressing catalytic domains is less reliable. A number of catalytic domains usually found in NRPS or PKS modules have been well-studied to allow a prediction of substrate specificities or the stereochemical outcome of the catalyzed reaction. However, it should be noted that only in rare cases complete core structures can be predicted whereas getting first hints on how a partial structure looks like or what a special domain does is frequently achieved. Aside of being subject to wrong predictions attributed to inaccurate models, the prediction of a biosynthetic product is additionally complicated by unexpected effects of tailoring enzymes and deviations from so-called textbook biosynthetic logic. As the latter is a common finding, ambitions to accurately predict the final product of a biosynthetic gene cluster is far from reliable or even unreasonable.

This thesis utilizes genome mining techniques to characterize and subsequently analyze biosynthetic domains in a detailed manner. In particular, the biosynthetic machinery of pellasoren and microsclerodermin production was identified on the basis of the *Sorangium cellulosum* So ce38 genome followed by an in-depth analysis of the involved catalytic domains. In the following, the capabilities and limitations of *in silico* predictions are exemplified for pellasoren biosynthesis.

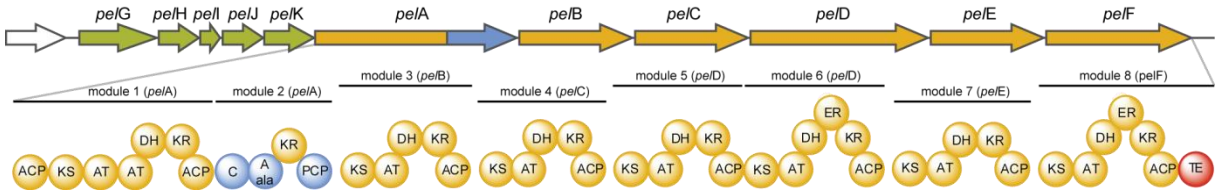


Figure 35: Pellasoren biosynthetic gene cluster as predicted by the antiSMASH2.0 annotation pipeline.[1] *All domains were correctly annotated. Based on this result a draft molecule can be predicted and manually curated as discussed in the text.*

Pellasoren is a linear NRPS/PKS hybrid natural product isolated from *S. cellulosum* So ce38. Analyzing the pellasoren gene cluster using the AntiSMASH2 annotation pipeline results in a set of biosynthetic genes that could largely fit the pellasoren structure as elucidated by NMR (Figure 35).[1] Pellasoren biosynthesis follows the colinearity rule, i.e. each module holds responsible for a single NRPS or PKS-related extension cycle. This type of textbook biosynthetic logic clearly facilitates prediction of the final structure which should in turn result in an acceptable structure proposal.

Although having good similarity, the predicted core structure looks different to the real molecule (Figure 36, A and D). This initial prediction is based on the number of modules found in the cluster and their presumed substrate specificity as identified by protein sequence alignments for the domains responsible for substrate discrimination. Similar analyses for the remaining catalytic domains reveal whether these are active and whether they can act in the course of biosynthesis. In

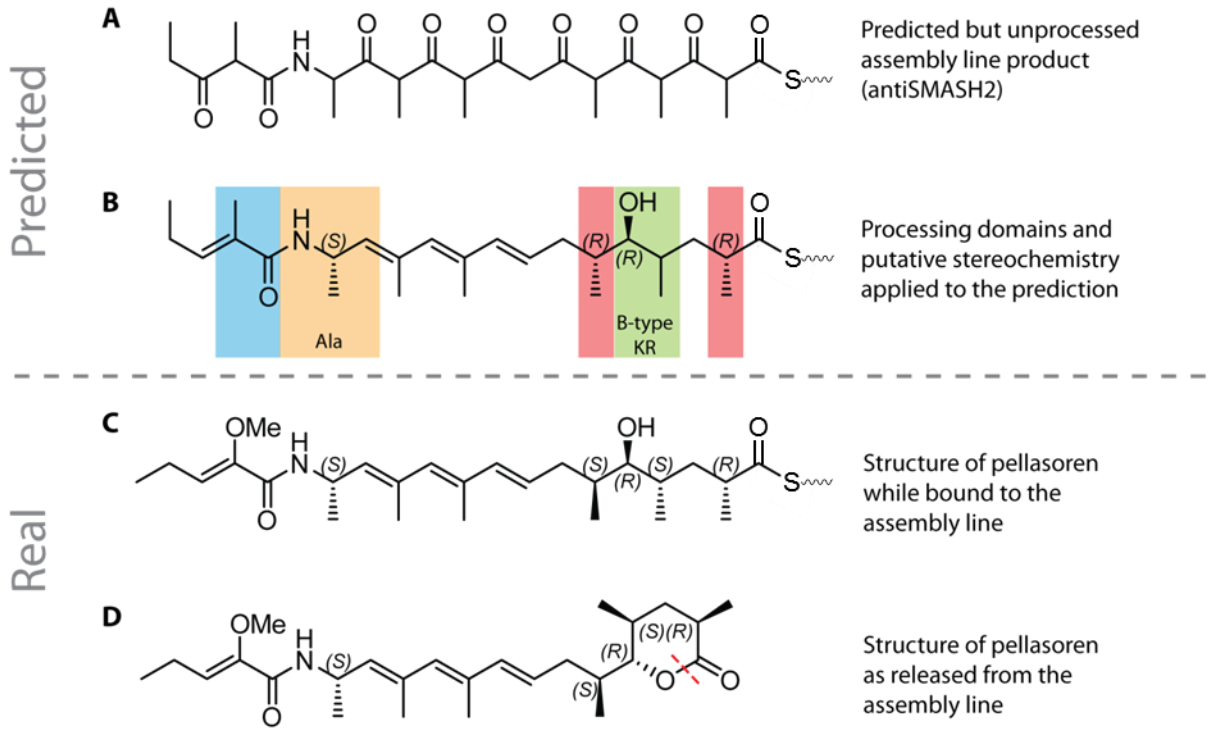


Figure 36: Comparison of predicted and real structure of pellasoren A. The colored areas mark four different aspects of in silico prediction: 1) the incorporation of uncommon extender units (blue, PKS part of PelA), 2) the choice between R and S-configured amino acids (orange, NRPS module of PelA), 3) the stereochemistry of enoyl reduction (red, PelD and PelF), and 4) the stereochemistry of ketoreductase domains (green, PelE).

the last step, the prediction of stereogenic centers is achieved by alignment of protein sequences of relevant domains in order to identify class-specific fingerprint motifs. The resulting structure shown in Figure 36 B is in very good agreement with the respective real structure as shown in Figure 36 C. However, the incorporation of the methoxy group (highlighted in blue) as well as the configuration of one methyl group (highlighted in red) could not be correctly predicted which in turn causes the structure to be slightly different to the real one shown in Figure 36 C. Alongside with this, the configuration of an alanine-derived stereogenic center has to be carefully reviewed (highlighted in orange). For that reason, the capabilities and limitations of the *in silico*-based methods related to the prediction of these specific structural moieties are discussed in the upcoming sections. They correspond to the highlighted groups shown in Figure 36 B and comprise an insight on the A-domain specificity toward D-amino acids, the incorporation of uncommon methoxymalonyl extender units by PKS modules, and the stereochemical outcome of KR and ER domains.

5.1.1 Incorporation of Amino Acid Building Blocks

One methyl moiety of the pellasoren scaffold is derived from an alanine that is incorporated by the NRPS module of the biosynthesis cluster (Figure 36, highlighted in orange). In this special case, it was not possible to determine the stereochemistry of the carbon atom by means of spectroscopic analysis or chemical degradation experiments. Hence, *in silico* analysis of the corresponding NRPS module provides an alternative approach to get further insights into the stereochemistry that is presumably set during biosynthesis of this particular structural unit. A typical minimal NRPS module of A-C-PCP architecture usually incorporates L-configured amino acids rather than D-configured ones. Such an assumption is reasoned by the fact that L-amino acids are easily available due to primary metabolism and by the observation that D-amino acids are derived from L-amino acids by an epimerization domain within a NRPS module (A-C-E-PCP). However, in case of alanine there is a certain chance that D-alanine is present in the primary metabolome owing to its requirement for murein biosynthesis. Indeed, direct activation of D-alanine was recently reported for NRPS-related biosynthesis of fusaricidin and leinamycin.[2,3] Leinamycin is the first example where a D-amino acid is directly loaded onto a PCP domain to prime the NRPS biosynthesis. In fusaricidin biosynthesis the scaffold comprises four D-amino acids whereas three of them can be explained by the action of an epimerase within the module. The fourth module features only the A-C-PCP architecture although it holds responsible for D-alanine incorporation, and *in vitro* assays with the respective adenylation domain indeed confirmed that direct loading of D-alanine is taking place. In light of these results, the substrate specificity toward alanine enantiomers seems to be ambiguous and a conclusion drawn just based on the domain architecture is not suitable for alanine. Recent studies of the crystal structure of the D-alanine-D-alanyl carrier protein ligase (DltA) involved in lipoteichoic acid biosynthesis in gram positive bacteria revealed that D-ala adenylation domains are

distinguishable from those being specific for L-ala.[4] Although myxobacteria do not have this type of biosynthesis the activation of D-ala in murein-related or NRPS-related biosynthesis is comparable as the adenylation domains belong to the same family of enzymes.[4] In line with this result, D-ala-specific A domains were aligned with PelA-A from pellasoren biosynthesis and the A domains found in microsclerodermin biosynthesis (Figure 37). Yonus et al. reported that L-aa incorporating A domains do have an alanine or glycine at position 268 whereas a cysteine, an isoleucine, or an aspartate point toward activation of D-configured amino acids.[4] Based on this alignment and the reported classifiers it was possible to determine which amino acid enantiomer is presumably activated.

In case of microsclerodermin and pellasoren biosynthesis, all A domains use L-amino acids which is in agreement with the alignment as shown in Figure 37. The stereogenic center of lipothiazole's cysteine is lost during thiazole formation but the above shown alignment suggests activation of L-cysteine. The microsclerodermin structure features two amino acid residues with *R*-configuration, a tryptophan and a pyrrolidone moiety which is derived from an asparagine. The *R*-tryptophan is based on an epimerization domain in the respective module and thereby represents the common way of how this particular stereochemistry is set on an assembly line. The configuration of the asparagine residue is *S*-configured based on the 268G found in MscF-A1 which is in agreement with the incorporation of fully labeled *S*-asn as verified by feeding experiments. However, it turned out that this stereogenic center is of *R*-configuration in the final microsclerodermin structure.

A domain					Pos. 268					Predicted confg.	Observed confg.	Compound
DltA_B.subtilis (ala)	T	F	M	F	**C**	G	E	V	L	D	D	-
FusA_A6 (ala)	Y	L	F	A	**C**	G	E	T	L	D	D	Fusaricidine
LnmQ (ala)	H	T	V	F	**C**	G	E	P	L	D	D	Leinamycin
PelA-A1_ala	A	L	N	L	A	G	E	A	L	L	L	Pellasoren
MscH-A1_gly_Soce38	K	I	L	C	G	G	E	A	L	L	-	Microsclerodermin
MscH-A3_gly_Soce38	R	A	L	C	G	G	E	A	L	L	-	Microsclerodermin
MscI-A1_gly_Soce38	R	V	L	C	G	G	E	A	L	L	-	Microsclerodermin
MscH-A2_trp_Soce38	A	L	V	V	G	G	E	S	C	L	L	Microsclerodermin
MscF-A1_asn_Soce38	A	L	I	L	G	G	E	A	L	L	L	Microsclerodermin
LipA-A1_cys	R	A	L	A	G	G	E	A	L	L	-	Lipothiazole

268G or 268A for L-aa

268C, 268I, 268D found for D-aa

Figure 37: Alignment of the D-alanine-D-alanyl carrier protein ligase (DltA) to known D-ala activating A domains from fusaricidin and leinamycin biosynthesis and the A domains as found in pellasoren, microsclerodermin, and lipothiazole biosynthesis. Position 268 is a bulky cysteine, aspartate, or Isoleucine for D-ala specific A domains. L-specific domains feature a glycine or alanine at this position. Numbering according to DltA.[4] Accession numbers for the sequences are 3E7W (PDB database, DltA), EF451155 (GenBank, fusaricidin, FusA), AF484556 (GenBank, leinamycin, LnmQ), HE616533 (EMBL, pellasoren, PelA), and KF657738 (GenBank, microsclerodermin, MscF, MscH, MscI)

Thus, there must be a configuration swap during biosynthesis, most likely as a result of the cyclization of the asparagine side chain. This finding illustrates well the limits of *in silico* prediction which are related to the presence of uncommon tailoring enzymes and deviations from textbook biosynthetic logic. After all, a hypothesis based on *in silico* analysis still needs to be verified by suitable methods such as full structural elucidation, total synthesis, or chemical degradation experiments. The myxobacterial compound classes characterized in this work agree well with the prediction of the amino acid-related stereogenic centers. Notably, all stereogenic centers were confirmed by degradation experiments in case of microsclerodermin or by total synthesis in case of pellasoren. Such information on biosynthetic domains in combination with a proven structural outcome of the respective domains is essential to further strengthen the predictive power of the underlying model.

After all, the direct loading of D-amino acids – at least for D-ala – seems predictable although the number of such adenylation domains is still rather low. The fact that this model is built on crystallographic data together with the highly conserved structure of adenylation domains makes it a valuable contribution to future computational tools.

5.1.2 The Glycolate Extender Unit

Pellasoren biosynthesis features an enol ether moiety, which is still rather uncommon amongst the products of natural product biosynthetic pathways. With regard to common biosynthetic logic of PKS systems, one would assume that a methoxymalonyl-CoA extender unit is loaded to the ACP by an AT domain in the same way like it is done for malonyl and methylmalonyl-CoA. However, this is not the case and it has been shown that the *N*-acetylcysteamine thioester of methoxymalonate or hydroxymalonate is not incorporated although it would be the biosynthetic mimicry for the involvement of methoxymalonyl-CoA in biosynthesis.[5] Instead, it has been hypothesized that the methoxymalonyl-moiety is derived from a glycolate attached to an ACP via a thioester bond with subsequent oxidation and methylation to yield methoxymalonyl.[6] This hypothesis is based on the first “glycolate extender" gene cassette which has been described for FK520 biosynthesis. This cassette covers five genes, one acyl carrier protein (FkbJ), one of unknown function (FkbH), one methyl transferase (FkbG), and two homologues of a fatty acid acyl-CoA dehydrogenase (FkbK and FkbI).[6] Later, the crystal structure of the dehydrogenase FkbI revealed that it acts on ACP-bound substrates rather than on CoA-bound ones.[7] In particular, FkbI catalyzes the second oxidation of a glycerate-derived template that is bound to FkbJ to yield the aldehyde as shown in Figure 38. In addition it has been proven by feeding studies with labelled glycerol and glycerate that the substrate is indeed glycerate albeit it is not known if the glycerate is phosphorylated.[8]

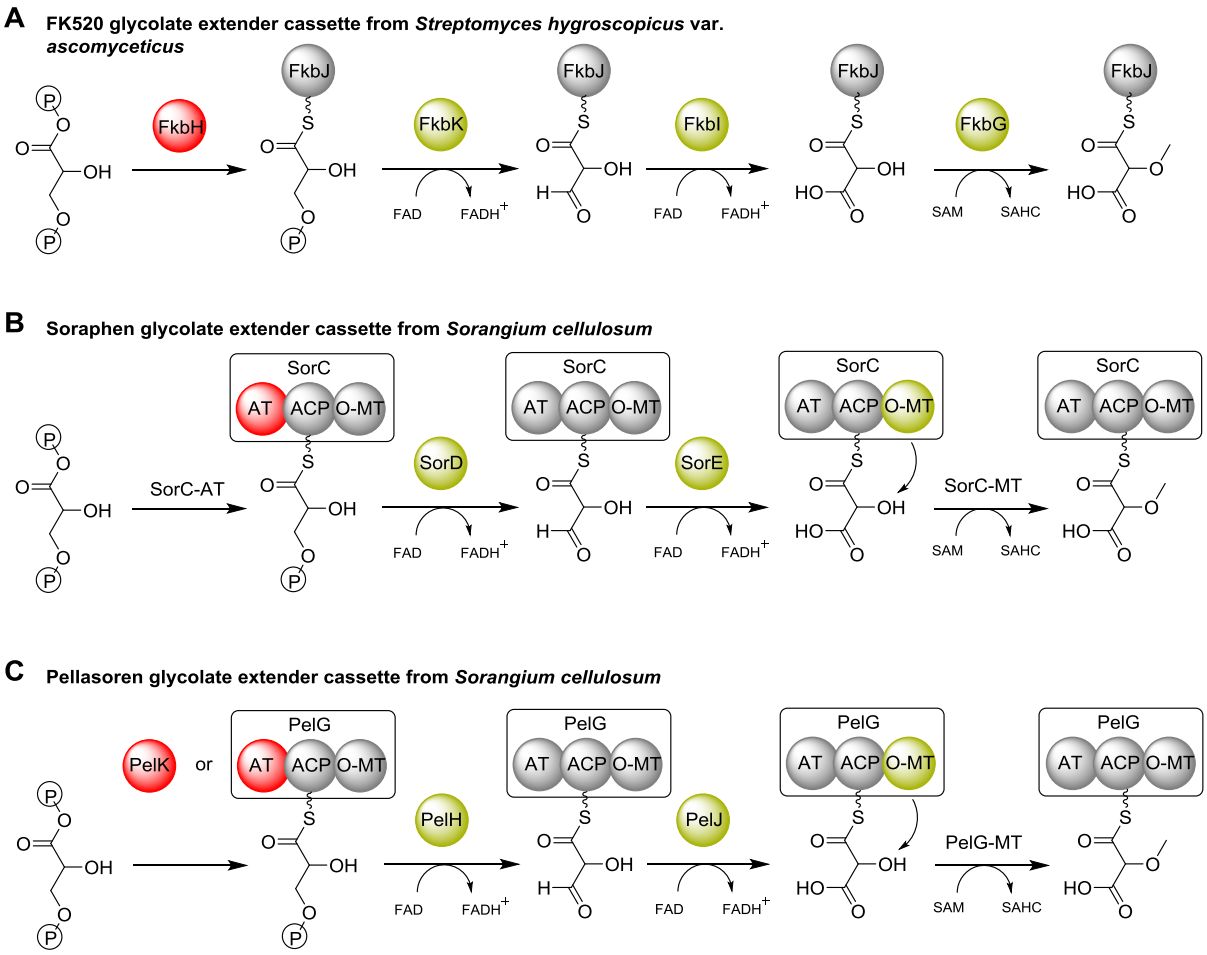


Figure 38: Proposed mechanisms underlying the biogenesis of the glycolate extender unit as found for three different biosynthetic pathways. ***A*** *The first gene cassette for glycolate extender biogenesis identified in a* Streptomyces *sp. involved in FK520 biosynthesis. The protein of unknown function FkbH is supposed to load bisphosphoglycerate onto the ACP domain.* ***B*** *Soraphen biosynthesis has no FkbH-like enzymes but rather uses an acyl transferase domain.* ***C*** *For pellasoren biosynthesis enzymes of both types are found which seems to be redundant.*

It has been shown that two different enzymes can load the glycerate in the very beginning of biogenesis.[6,8] For FK520 biosynthesis there is an enzyme of unknown function named FkbH likely loading the glycerate starter while in soraphen biosynthesis no homologue of FkbH is found. Instead an additional acyl transferase (SorC-AT) is most likely catalyzing the loading step (Figure 38). Intriguingly, pellasoren biosynthesis has both an AT domain (PelG-AT) and an FkbH-like protein (PelK). This fuels speculation whether there are two congruent ways of loading the same substrate, or whether these enzymes have preferences toward slightly different substrates, e.g. 3-P-glycerate or 1,3-P_2-glycerate.

In silico analysis of the So ce38 genome led to correct annotation of the single domains belonging to the glycolate gene cassette. While this indicates the strain's capability to produce glycerate-related extender units it is currently impossible to infer which module of the biosynthetic assembly line incorporates this special extender unit. In theory, the pellasoren scaffold could have a methoxymalonyl unit incorporated at all PKS-related positions or even multiple times. One possible approach to propose a module incorporating an uncommon extender unit is based on the phylogenetic analysis of KS

domains as they have to accept the extender unit to catalyze the elongation step. Such a phylogenetic analysis is based on alignments conducted on protein level. For pellasoren biosynthesis, PelD-KS2 and PelA-KS1 are different compared to the other KS domains found in this cluster. PelD-KS2 is supposed to catalyze bond formation between a malonyl precursor and a methylmalonyl-ACP which likely causes the "difference". PelA-KS1 might be different because of its priming function or because of conducting a methoxymalonyl substrate. Although this is not a clear-cut result, it points toward incorporation of this special extender unit by module 1, which would be correct in case of pellasoren.

Interestingly, microsclerodermin biosynthesis incorporates hydroxymalonyl extender units in the early stages of biosynthesis although there is no glycolate extender gene cassette in proximity to the cluster. This finding could be rationalized by "pellasoren's" glycolate extender gene cassette being responsible for both the pellasoren and microsclerodermin pathway in So ce38. It has indeed been reported that the glycolate extender cassette is not necessarily located in proximity to the secondary metabolite gene cluster.[9] However, the dual use of the cassette remains puzzling as pellasoren biosynthesis solely incorporates a methylated glycolate extender while microsclerodermin biosynthesis uses a non-methylated one. The fact that the methyl transferase is translationally linked to the ACP harboring the extender unit supports the assumption that a methylation happens during maturation of the methoxymalonyl extender unit. This is obviously not necessarily the case as this would prevent the biogenesis of hydroxymalonate for microsclerodermin biosynthesis. Searching the genome of *S. cellulosum* So ce38 for an alternative glycolate cassette was not successful; hence, the extender cassette has to deliver both methoxymalonyl and hydroxymalonyl extender units. This brings up the discussion at which point of biogenesis the *O*-methylation takes place. The trifunctional protein as found in soraphen biosynthesis and in So ce38 would imply to have the methylation at an early stage.[8] However, the fact that two different precursors are generated in So ce38 – and indeed used in metabolite biosynthesis – supports a late methylation step (Figure 38). The discrimination between both extenders is then most likely achieved by the KS domains of the respective PKS modules PelA-KS1 and MscC-KS1.

Inactivation of the glycolate extender cassette could lead to new insights regarding the use of this special extender unit in *S. cellulosum* So ce38. The expected result of this single knock out would be the abolishment of pellasoren and microsclerodermin production at the same time. Since the cassette is part of the pellasoren operon, a knock out will likely affect the expression of the remaining pellasoren cluster owing to downstream polar effects. However, the microsclerodermin operon would remain intact and thus disruption of the glycolate cassette will either result in abolishment of microsclerodermin production or in new microsclerodermin derivatives based on a common extender unit such as malonyl-CoA.

5.1.3 The Stereochemistry of Enoyl Reduction in PKS

The criteria that govern the configuration of the stereogenic center being set during enoyl reduction of α-substituted extender units has to date only been poorly investigated. Kwan et al. set up a model to predict the stereochemical outcome based on an alignment with numerous ER domains.[10] The study is based on 39 domains from *Streptomyces* sp., 3 from *Sorangium* sp., and 12 from *Mycobacterium* sp. alongside with several others. There is a certain chance that the resulting model is biased toward *Streptomyces*-related ER domains; however, the result should be an adequate basis for myxobacterial domains as well. As a result of their study, Kwan et al. highlighted a tyrosine (Tyr52) that is highly conserved amongst ER domains yielding a 2*S*-configured stereogenic center when the extender unit is bound to the ACP. This Tyr52 residue most likely discriminates which stereochemistry is created during reduction by NADPH. By mapping two ER domains onto the crystal structure of the *E. coli* quinone oxidoreductase QOR (PDB ID 1QOR) they eventually proposed that Tyr52 is within the active site pocket and thereby reduces the space that is available in the pocket (Figure 39 C).[10] The involvement of Tyr52 was verified by site directed mutagenesis with subsequent analysis of the respective mutant's biosynthetic products. However, results of these experiments already indicated that Tyr52 is not exclusively directing the stereochemical outcome.

For pellasoren biosynthesis, an ER domain classification based on Tyr52 does not hold true for one of the two ER domains, PelD-ER1. Both domains are supposed to create 2*R*-configured carbons based on alignment of the respective ER domains. However, total synthesis of the pellasoren molecule has unambiguously proven that one ER yields a 2*R*-configuration while the other ER domain creates a 2*S*-configuration (Figure 39 B). Similarly, the proposed Tyr52 classifier does not apply for one of the ER domains found in soraphen biosynthesis as characterized in *Sorangium cellulosum* (SorA-ER3).[11] Hence, out of 5 domains from *Sorangium* sp. analyzed with respect to the method Kwan et al. proposed, two do not match. In both cases a Leu52 is present while the structure has *S*-configuration at the respective position. For soraphen biosynthesis this may be reasoned by the methoxymalonyl extender unit. However, in pellasoren biosynthesis a typical methylmalonyl-CoA extender is used. Since the genome of *S. cellulosum* So ce38 was sequenced four times independently, the chance to have a sequencing error is negligible as well. Remarkably, both ER domains of the pellasoren cluster set a different stereochemistry while being identical in 291 out of 316 amino acids (92.4 %). Based on the protein structure of 1QOR, the differences between both Pel-ER domains were located within the 3D structure and it turned out, that the only obvious difference within the active site groove is Arg43 (Figure 39 C, numbers according to 1QOR).

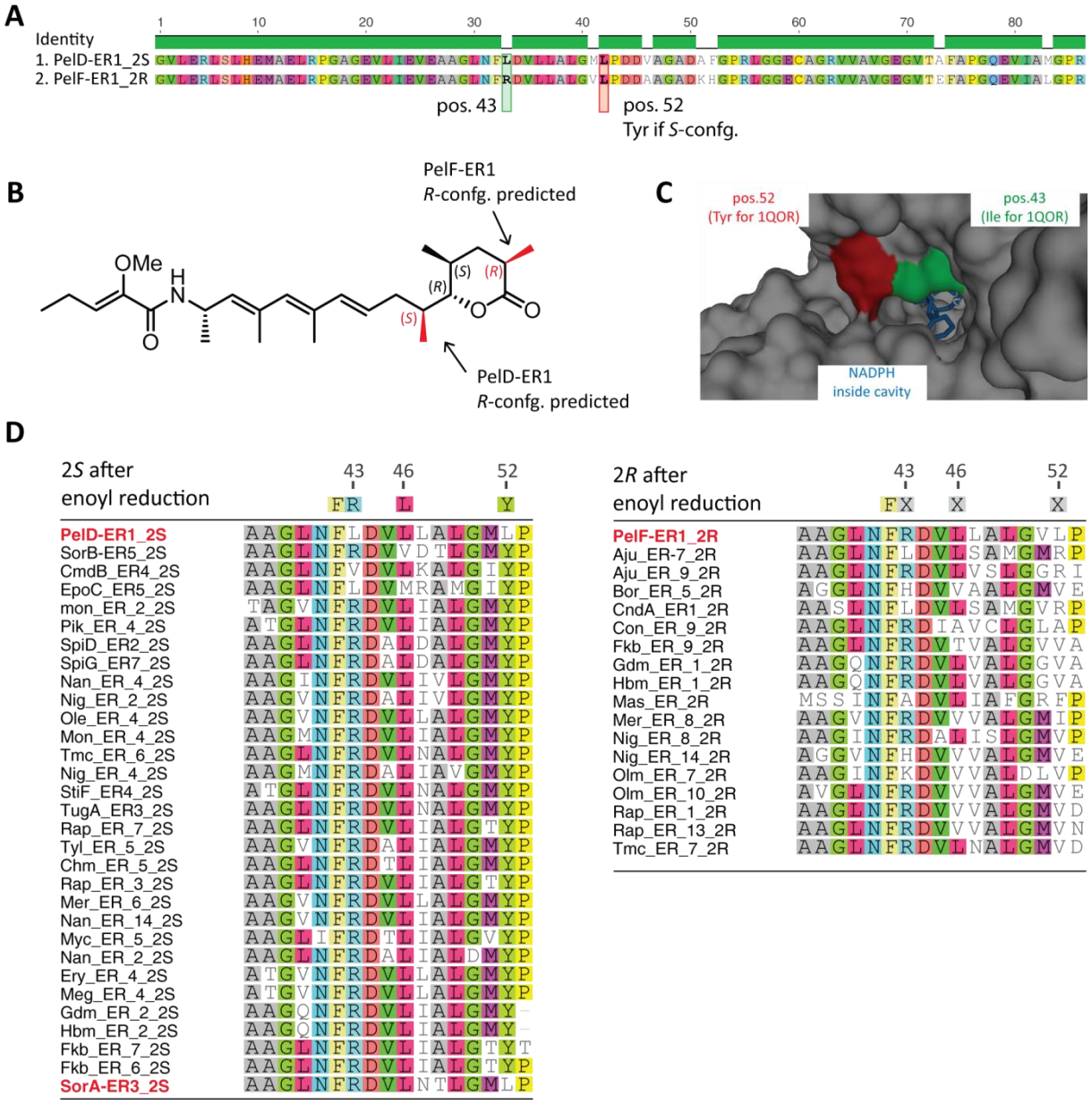


*Figure 39: **A** An alignment of both Pel-ER domains shows remarkable identity although different stereogenic centers are set during biosynthesis. Marked positions correspond to the numbering of the 1QOR sequence. **B** Pellasoren structure: Mismatch between observed and predicted stereochemistry. **C** Crystal structure of 1QOR with residues 43 and 52 highlighted. Both residues directly affect the size of the groove and likely undergo interactions with the substrate thereby influencing the stereochemical outcome of enoyl reduction. The cofactor NADPH is shown in blue. **D** Alignment of the active site region of several ER domains that create 2S or 2R stereochemistry, respectively. 2S-ER domains show a higher degree in conserved residues compared to 2R domains. The two domains of pellasoren and soraphen biosynthesis that produce 2S stereogenic centers although lacking the presumed conserved residue Tyr52 are highlighted in red.*

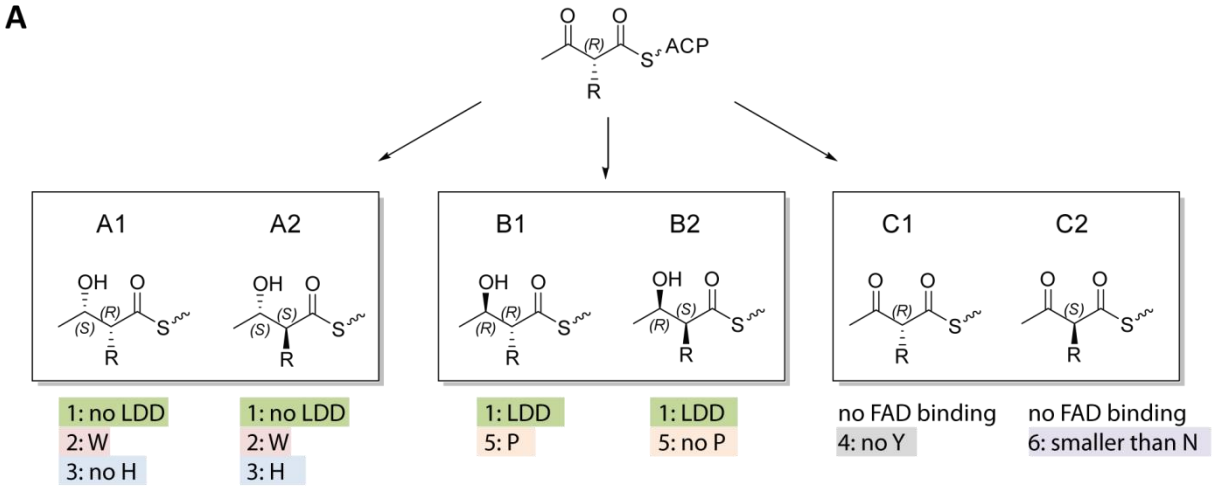
Both Arg43 and Tyr52 belong to the same α-helix and both are in contact to the substrate and the cofactor which could be sufficient to influence the stereochemical course of the reduction. Its position within the 1QOR structure underpins this hypothesis. By comparing the conserved residues within the active site groove it becomes evident that ER domains generating a 2S stereochemistry do have the combination Phe42, Arg43, Leu46, Tyr52 (Figure 39 D). In particular, Phe42 points towards the cofactor while the basic Arg43 points into the substrate cavity. For those domains that hold responsible for 2R

configuration, positions 43 and 46 are frequently different in addition to the lack of Tyr52. The combination of these residues likely governs the stereochemistry as the corresponding helix directly affects the cavity size and thus protein-substrate interactions. Aside of this, the origin of the divergent stereochemistry in pellasoren may be related to some effect induced by the tethered substrate and not solely governed by the amino acid sequence of the ER domain. A dramatic influence of the substrate on the stereochemical outcome was recently reported for KR domains.[12,13] This can be attributed to different ways of how a substrate approaches the active site groove.[14] ER domains are likely prone to similar effects.

After all, the underlying model by Kwan et al. needs to be improved by including ER domains of, for example, myxobacterial origin which will most likely result in a better model. At least the key amino acid residues that are supposed to classify a 2S stereochemistry should be extended to the combination Phe42, Arg43, Leu46, Tyr52 and not rely solely on Tyr52. Moreover, the absence of Tyr52 is no guarantee for 2R stereochemistry but rather a hint. For a full understanding of stereochemical discrimination in enoyl reduction a crystal structure of PKS-related ER domain seems mandatory. With respect to their high identity on protein level and yet diverse stereochemical outcome, the ER domains of pellasoren biosynthesis are certainly a good choice to study enoyl reduction in more detail.

5.1.4 The stereochemistry of Ketoreduction

The principles that govern stereochemistry in keto reduction are heavily studied compared to enoyl reduction. Based on two crystal structures, Keatinge-Clay defined six different types of KR domains, four are based on the diverse configurations set at the extenders C2 and C3 atoms whereas two are related to KR domains that lack the reduction capability at C3, but isomerize C2 (Figure 40).[14] The fact that *R* and *S*-configured C2 atoms exist suggests an epimerization activity since the AT domains solely loads (2*S*)-methylmalonyl-CoA onto the ACP (at least there is no other type of AT domain known so far). The stereogenic center of methylmalonyl is inversed to 2*R*-configuration upon claisen condensation mediated by the KS domain; hence, 2*R* is the typical configuration set by a PKS minimal module. The outcome of keto reduction for the different KR types is depicted in Figure 40 A. At this point it is important to mention, that the stereochemical classification of KR types is only correct as long as the chain tail has per definition a lower priority than the chain head in terms of Cahn-Ingold-Prelog nomenclature. Moreover, double bond configuration may be predicted based on the KR domains as the stereogenic center of the KR-derived 3-hydroxy ester is directing the succeeding dehydration toward a (*Z*) or an (*E*)-double bond, respectively. The typical (3*R*)-hydroxy carbonyl moiety tethered to ACP via an thioester bond will result in an (*E*)-double bond.[15]



B

Fingerprint position	loop 1	catalytic region 2	3	4	5	6	predicted KR type	real structure	structural element
MscA-KR2_MSr9139	H L A G S L D D G V L	S S I A S I L	G S A G Q G N	Y	A	A S A	B1	(B)	*E*-double bond
MscA-KR2_soce38	H L A G A L D D G V L	S S I A S V L	G S A A Q G N	Y	A	A N A	B1	(B)	*E*-double bond
MscC-KR_MSr9139	H A A G V G G Y C E L	S S I A S L	W G S R G Q A H	Y	A	A N A	A1	A1	2*R*,3*S*-confg.
MscC-KR_Soce38	H A A G V A G A R D L	S S I A S L	W G S R G Q A H	Y	A	A N A	A1	A1	2*R*,3*S*-confg.
MscG-KR_MSr9139	Y A A A I - D D A G L	S S L S A I L	G G R G L A A	Y	A	C H L	B1	-	pyrrolidone
MscG-KR_Soce38	Y A A G V - D D P A L	S S L S A I L	G G R G L G A	Y	A	C S L	B1	-	pyrrolidone
MscI-KR_MSr9139	H A A G A R G D G T F	S S T S A I L	G G L G L G P	Y	A	A N R	(A1)	A	3*S*-confg.
MscI-KR_Soce38	H A A G A R G D G T F	S S T S A I L	G G L G L G P	Y	A	A N R	(A1)	A	3*S*-confg.
PelA-KR1	H A A A V L D D H T L	S S A L S L F	G S P G Q A N	Y	T	A S N A	B1	(B)	*E*-double bond
PelB-KR	H A A A V L D D H T L	S S G A S L F	G S P G Q G N	Y	A	A S A	B1	(B)	*E*-double bond
PelC-KR	H A A A V L D D H T L	S S A S A L F	G A P G Q G N	Y	T	A A N A	B1	(B)	*E*-double bond
PelD-KR1	H A A G V L D D G V L	S S L S S V L	G S G G Q G N	Y	A	A A N A	B1	(B)	*E*-double bond
PelD-KR2	H A A A V L D D H T L	S S A A S L F	G S P G Q G N	Y	A	A A N A	B1	-	fully reduced
PelE-KR	H T A V N F D N H P V	S S V S V L L	G L P G L G N	Y	L	A A N A	(B1)	B1	2*R*,3*R*-confg.
PelF-KR	H A A A V L D D H T L	S S A S A L F	G S P G Q G N	Y	A	A A N A	B1	-	fully reduced

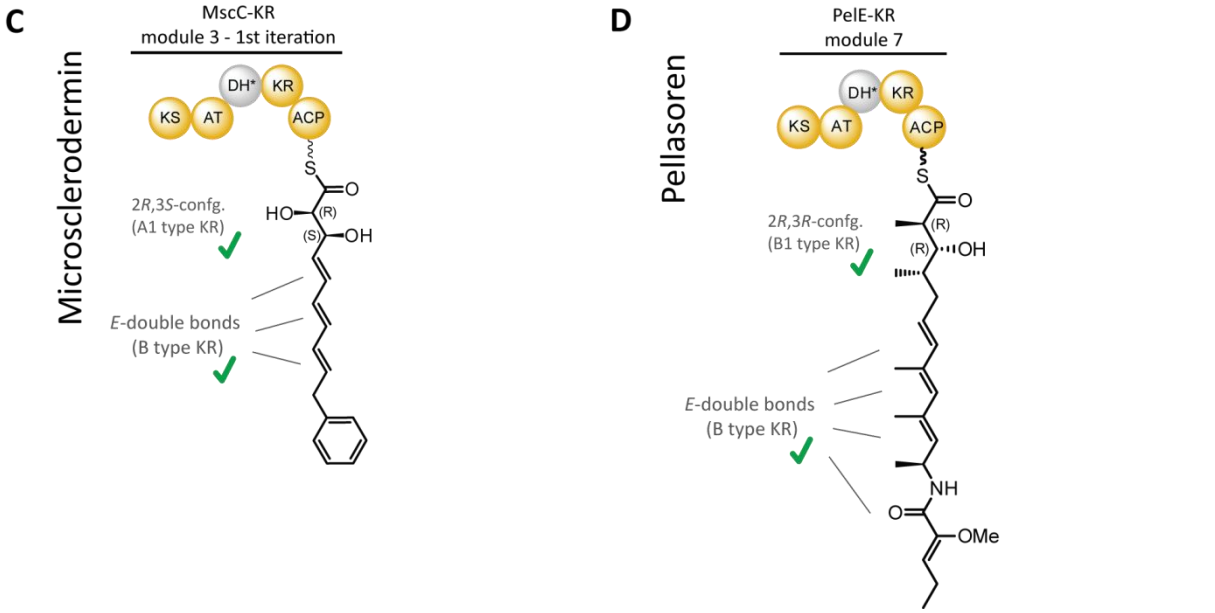


Figure 40: ***A*** *The stereochemistry of ketoreduction as found in PKS-related biosynthesis is predictable by checking for the presence or absence of fingerprint motifs within the KR protein sequence. Six types of KR domains exist.* ***B*** *Analysis of the fingerprint motifs for the KR domains of microsclerodermin and pellasoren biosynthesis.* ***C*** *The first iteration of microsclerodermin MscC-KR results in the same stereochemistry as predicted. The outcome of the second iteration cannot be determined as the 3-OH group is transformed to an amino group. The effect of this transformation on stereochemistry is not known.* ***D*** *Prediction of the stereochemistry set by pellasoren PelE-KR is in accordance to the real structure.*

With this knowledge, the structure of pellasoren and microsclerodermin is compared to the information derived from alignment analysis. In pellasoren biosynthesis, PelE-KR creates a 2*R*,3*R*-configured elongation product. In microsclerodermin, the iteratively acting MscC-KR creates a 2*R*,3*S*-configured product in the first iteration cycle and MscG-KR creates a 3*S*-hydroxyl group. The respective KR domains are aligned in order to compare the *in silico* prediction to the real structure and indeed all predictions fit to the observed stereochemistry (Figure 40 B). Also the double bond configuration fits to the predicted B-type KR domains.

Thus, all configurations found in pellasoren and microsclerodermin are in agreement with the *in silico* results. It can be concluded that prediction of KR domain specificity is reliable by referring to currently available classifiers. However, uncertainties may be introduced by variations of presumably conserved motifs, e.g. the LDD motif which has been regarded as highly conserved for B-type domains is replaced by a FDN motif in case of PelE-KR Figure 40. Such deviations of typical classifiers may complicate drawing conclusions from results of *in silico* analyses.

5.1.5 Conclusion

In silico based prediction of NRPS or PKS-related biosynthetic pathways is a powerful tool for natural product research. However, to comment on the power of this technique, a clear cut between (1) predicting a core structure of a molecule and (2) characterizing single domains of a biosynthetic pathway is essential.

Predicting a tentative structure correctly based only on *in silico* analysis is hard to accomplish. A first idea of how the core structure of a secondary metabolite would look like is in principal possible as exemplified for pellasoren. However, pellasoren biosynthesis follows the co-linearity rule and does not exhibit complicating factors. When complexity increases predictions are less reliable, e.g. many natural products are cyclized by a terminal TE domain, a process that is not yet linked to special structural features and cannot be foreseen. For example, the way microsclerodermin is cyclized (lactame formation) is not predictable by means of *in silico* analysis. Aside of this, biosynthetic pathways that do not follow common textbook biosynthetic logic are common. As soon as iterations occur or tailoring enzymes may act on the compound, the diversity of derivatives increases dramatically. The number of iterations is currently not predictable although hints on whether an iteration can occur may be identified based on alignment analysis,[16] e.g. in case of microsclerodermin an iterative PKS module is assumed based on the phylogenetic analysis of KS domains (Chapter 3). The position where tailoring enzymes act, and more important, how they act is also not known and remains a mere assumption when the structure is not yet known. A good example is the pyrrolidone moiety of microsclerodermin. The same holds true for trans-AT clusters that frequently show up with a “distorted” module organization. All these aspects are the reason why it is neither possible nor reasonable to create an *in*

silico model and use the resulting “molecule” to search for the resulting hypothetical *m/z* value in LC-MS data.

Nevertheless, various bioinformatics tools have become indispensable helpers when it comes to an in-depth characterization of single catalytic domains. All information gathered by such analyses eventually supports the overall process of full structural elucidation, especially in terms of characterizing stereogenic centers, which may be challenging when dealing with natural products. *In silico* models may assist this process although it was shown that several types of analyses not yet reached high reliability as exemplified for the ER domains. At the same time, predictions of amino acid-derived stereogenic centers or those related to KR domain activity are trustworthy.

Irrespective of the approach that is followed with these techniques, a full characterization by suitable techniques such as NMR is inevitable whereas models to predict biosynthetic specificities are powerful tools to support this process. With more and more biosyntheses characterized, *in silico* models will gain reliability which may in turn lead to higher success rates in assignment of preferred substrates and stereochemical outcome of single domains. Moreover, and even of higher importance, is the understanding of catalytic reactions when it comes to genetic engineering aiming to optimize existing biosynthetic pathways or even create new ones using synthetic biology. All these methods benefit of each other by contributing to models that aim to fully understand how biosynthesis works down to the last detail.

5.2 Secondary Metabolomics

While the current capabilities and limitations of genome-based *in silico* analysis of biosynthetic pathways were extensively discussed in the preceding sections, this chapter addresses the analysis of biosynthesis products by using LC-MS and advanced methods for data mining with complex metabolite-containing extracts.

LC-MS has evolved as the most widely applied method for the analyses of complex small-molecule mixtures as it combines two orthogonal and powerful yet robust separation techniques at a time. While chromatographic separation results in compound-specific retention based on physicochemical properties such as molecule polarity, separation according to *m/z* ratios and subsequent acquisition of fragment spectra yield a plethora of highly informational data on distinct compounds. However, both the tools to effectively mine these data and the development of analytical methods to produce suitable data as input for computational tools currently still lag behind the standards in the field of proteomics, which has benefitted over years from the developments of LC-MS techniques and vice versa. In proteomics, highly efficient chromatographic separations in combination with sensitive MS and MS/MS techniques constitute a crucial prerequisite for the analysis of digested proteins in whole cell lysates.[17]

While the computational analysis of mass spectral data from peptides has reached a highly automated and at the same time impressively reliable level, this is to date not the case for any other type of small molecules such as secondary metabolites. Nevertheless, those established LC-MS techniques are advantageous for all analytical problems where complex biological samples need to be identified in a fast and robust fashion in order to get detailed insights into the sample's molecular composition. With respect to natural product research, this aim becomes even more important as research rapidly developed during the past couple of years away from the traditional analysis of single compounds from individual strains toward large-scale screening projects in steadily growing collections. To achieve a high analytical depth per sample – including the identification of both known and unknown compounds – and at the same time enabling an overview across numerous samples is the primary objective of secondary metabolomics.

Working with microbial producers of secondary metabolites involves dealing with crude extracts of cells or culture supernatants being rich in diverse chemical classes of small molecules including the secondary metabolites of interest. With respect to the amount and complexity of mass spectral data produced in today's natural product screening workflows it becomes evident that automated data processing is needed to extract the necessary information from the acquired data in order to achieve dereplication and maintain at the same time a comprehensive overview of metabolite profiles.[18]

The identification of known compounds in samples using LC-MS is sufficiently well accomplished by database queries based on matching retention time, accurate *m/z,* isotope pattern analysis or spectral matching of fragment data. In addition, these targeted methods may be readily included into largely automated screening workflows. However, the most challenging goal of secondary metabolomics is the identification of *new compounds,* thus demanding for new methods that specifically address the identification and classification of unknown analytes rather than searching only for knowns. For obvious reasons the principles of simple spectral matching are not applicable to identify unknowns since – as a matter of fact – no reference data is available.

This shortcoming has been recently addressed by a growing number of computational methods dealing with the characterization and identification of the *unknowns*, i.e. compounds that are not included in any database. When dealing with unknowns an important hallmark is the identification of a correct sum formula. Here, considering the isotope peak pattern is regarded as a highly efficient measure to narrow down the number of putative sum formulas that can be generated just based on an exact *m/z* value.[19,20] Computational tools such as Sirius[21], mzMINE[22] and Bruker's SigmaFit increase the positive hit rate by comparing the measured isotope pattern to the set of calculated ones followed by a scoring of hits. Another promising method of getting a correct sum formula takes fragment spectra into account. The Sirius2 software[21] and Bruker's SmartFormula3D™ implement this approach which is based on the constraint that a fragment's sum formula has to be a subset of the parent's sum formula. Such a

filtering reduces the false-positive rate by excluding parent sum formulas that could never match the fragment ones; however, it should be noted that accurate m/z values of fragment ions are mandatory for the success of this method.

Besides supporting sum formula determination, it is well-known that fragmentation data can essentially aid to generate detailed structural information of unknown compounds.[18,23] Especially electron impact (EI) ionization as commonly used in GC-MS setups is a well-established and well-studied method to identify and characterize compounds as it causes a strong, reproducible, and highly informational fragmentation of analytes.[24] In a similar fashion, collision induced dissociation (CID) as frequently used in LC-ESI-MS setups provides insight into a compound's structure by creating fragments. While a detailed evaluation of MS/MS or MS^n data used to rely on the expertise of a skilled user, such a manual approach is inappropriate when considering the flood of data that is created for a single LC-MS/MS run of a complex sample; as an example, each auto-MS^2 run executed as part of this thesis frequently covered around 1000 MS/MS spectra. Hence, software-aided methods are urgently needed to generate useful information from fragment spectra in a more or less unsupervised fashion.[18,23] With respect to this, several objectives and methods can be distinguished:

- Methods for clustering of MS/MS data according to similarity
- Methods that classify MS/MS spectra by scoring their similarity to known compounds
- Methods that match MS/MS-derived information to biosynthetic gene cluster
- Methods aiming at *de novo* prediction of compound structures based on MS/MS data

Since a comprehensive analysis is the ultimate goal in secondary metabolite-related screening projects a totally unbiased approach of data mining is highly appreciated. In line with this, unsupervised clustering of MS/MS spectra is a powerful method to classify spectra according to similarity, which is in turn advantageous to identify structurally related compounds, e.g. members of a compound class. Yang et al. described a method of "molecular networking" which is capable to link MS/MS spectra that are most likely related to each other as indicated by similar fragmentation.[25] Moreover, it has been shown that this technique has the potential to identify presumably new derivatives since those tend to cluster with known compounds of a given class. For example, spotting formerly unknown spectra together with a spectrum of a known compound triggers further work on the compounds represented by these new spectra. Böcker et al. introduced a different approach which creates hypothetical fragmentation trees for signals observed in fragment spectra.[26] The method uses fragment spectra with accurate *m/z* information and evaluates which fragment signals are related to each other based on neutral loss analysis. With the knowledge about typical neutral losses and rudimental chemical logic the method creates a fragmentation pathway for the signals found in a MS/MS spectrum of one parent ion.[27] In a

second step, a grouping of distinct MS/MS spectra can be achieved by alignment of the obtained fragmentation trees. Both spectral networking and fragmentation tree alignment may be used for clustering unknowns as well as for a similarity-based matching of unknowns to known compounds. The power of the spectral networks approach has been exemplified by analyzing microbial cultures using imaging mass spectrometry. [28,29]

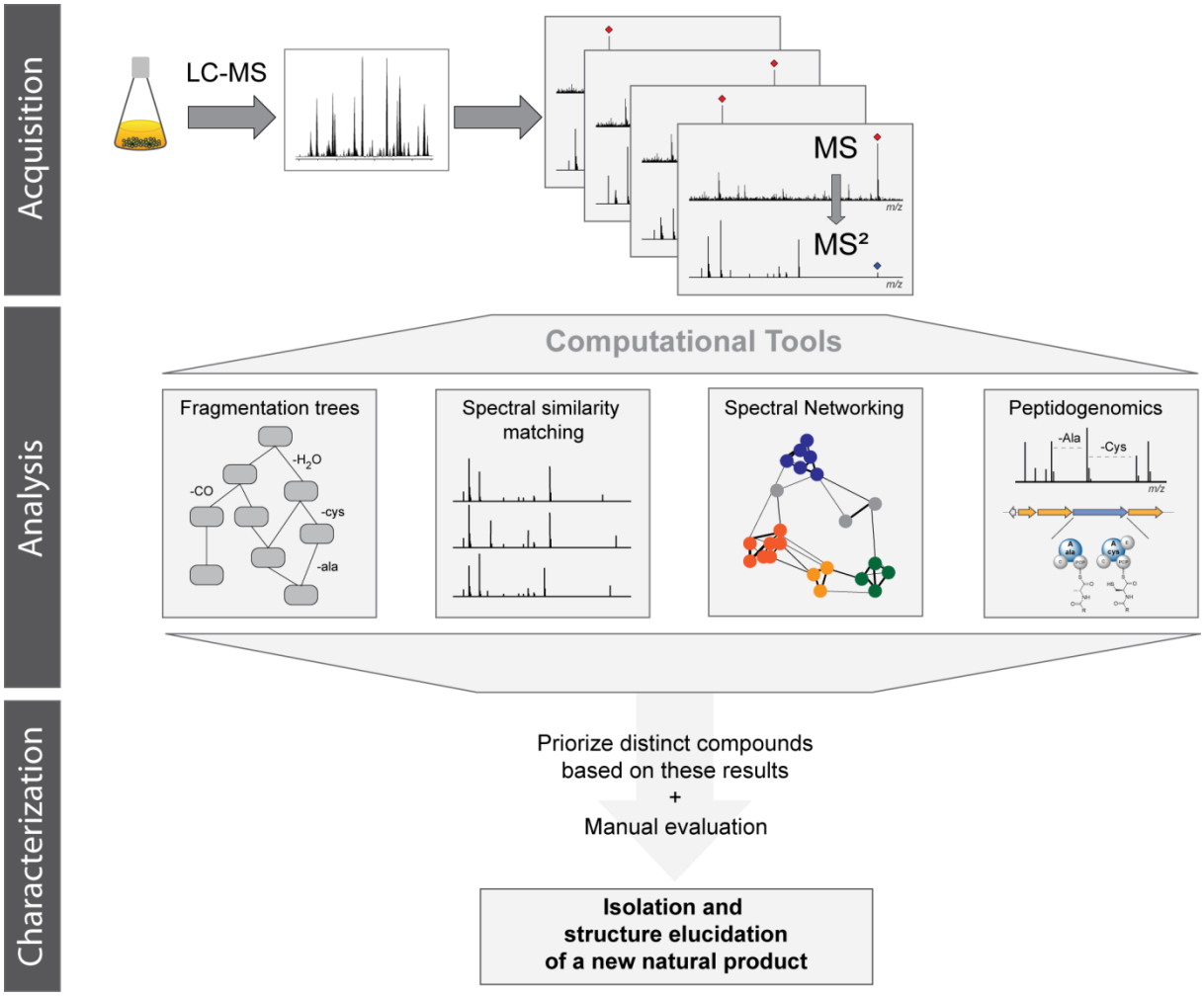


Figure 41: Possible secondary metabolomics workflow benefitting from computational tools that evaluate LC-MS/MS data. Acquired fragment spectra are essential for these tools and allow for a structural classification of unknown compounds at an early stage of natural product research. A pre-selection of precursor ions as introduced in this work leads to an increased informational content of the raw data that is used by subsequent data mining tools.

Other methods of data mining critically depend on the availability of reference spectra in order to perform a comparison on spectral level, e.g. by using the similarity of fragment spectra[30] or by establishing a computational model by machine learning algorithms applied to known spectra in a first pass.[31] Kersten et al. showed that indicative fragmentation patterns can lead to the identification of characteristic neutral losses attributed to amino acid building blocks or glycosylations upon which the genome is searched for matching biosynthetic modules. These techniques have been termed peptidogenomics and glycogenomics.[32,33] Ambitions to use MS/MS data to *de novo* predict a structure

are still limited to peptide-like molecules owing to their characteristic and well understood fragmentation. Here, notable progress has been recently made with the computational analysis of cyclic peptides, e.g. by the CYCLONE tool.[34,35] However, it must be noted that this limitation to peptide-like analytes will certainly lead to a biased result when analyzing complex samples. In summary, all these methods can contribute to comprehensive screening approaches in natural product research.

Despite their different underlying methods and scopes, the above-mentioned approaches have one thing in common: They rely on high-quality MS/MS spectra whereas accurate mass acquisition is beneficial or even mandatory for the success of analysis. As these methods are located at the very downstream end of mass spectrometric analysis workflow it is of utmost importance to take care of the preceding steps, namely the acquisition and preprocessing of MS/MS data. This is exactly where the method devised in this thesis acts.

A few publications specifically address the issue of improving the informational content of small-molecule MS/MS spectra or increasing the number of meaningful spectra obtained from a sample.[36] With respect to MS/MS precursor selection, a pre-selection of relevant ions may be achieved by scanning for special isotope ratios, e.g. when screening for isotopically labelled scavenger compounds like it has been shown for glutathion.[37] While this highly targeted MS/MS technique is capable to direct MS/MS fragmentation towards ions of interest, its application is limited to a single marker in one specialized context, and thus not useful for comprehensive screening. Other approaches are based on indicative neutral losses in MS^2 scans upon which the mass spectrometer will initiate a MS^3 scan.[38,39] Such neutral loss scan techniques lack high accurate mass acquisition as they are commonly conducted on low resolution but very fast QqQ-MS systems. Although being useful for a range of applications including proteomic investigations and biomarker studies, the main drawback of these methods is the very high bias introduced by the type data-dependent acquisition is achieved. Schmidt et al. reviewed approaches that deal with hypothesis-based selection of precursor ions in proteomics applications which they refer to as inclusion-list driven MS.[36] Those methods are following the idea of finding relevant features in a first pass followed by subsequent fragmentation of the features in a second step in order to improve peptide coverage in shot-gun proteomics.[40,41] Jaffe et al. introduced the PEPPeR workflow which specifically identifies and selects features related to differentially expressed proteins for subsequent fragmentation.[42]

In line with this, the combination of statistical data evaluation and SPL-MS^2 introduced in this work is a promising approach for natural product analytics as it is totally unbiased with respect to the precursor ion selection. The method is based solely on the comparison of sample-specific molecular feature sets which consist of retention time, *m/z*, and intensity. This comparison has no preference toward specific compound classes or structural moieties. In addition, the method devised in this thesis is suitable for the rather time-consuming acquisition of high accurate MS/MS spectra owing to the fact

that MS/MS acquisition time is exclusively devoted to measuring signals of interest. In comparison to auto-MS² methods using intensity-based precursor selection, acquiring MS/MS data in this new mode is beneficial for the identification of both known and unknown analytes for two reasons:

- The MS/MS coverage amongst potentially interesting compounds is tremendously increased owing to the fact that precursor signal intensities are irrelevant.
- The MS/MS data set is of high informational content as most part of the data is related to relevant signals.

Thus, when working with these improved MS/MS data sets, the results of computational methods applied downstream – e.g. unsupervised clustering using spectral networking or fragmentation tree construction – will exclusively perform on relevant data. This pre-filtered approach is considered highly advantageous, for example when thinking of complex cultivation media which will certainly add a plethora of matrix signals. Such nutrient molecules could potentially exhibit secondary metabolite-like structures and fragmentation behavior, e.g. peptides. Hence, a data set which covers MS/MS spectra of matrix signals is prone to introduce a bias to the clustering algorithm as these “irrelevant” signals could easily outnumber the compounds of interest. For example, a NRPS-derived natural product with several proteinogenic amino acid building blocks likely cluster with media-derived peptides of similar constitution. This potential influence on a computational analysis workflow is easily avoided by pre-filtering the MS/MS precursors as proposed in the approach presented in this work.

In conclusion, carefully selecting the input data is instrumental for increasing the significance of results from unsupervised computational analyses, thereby improving the utility of such tools for natural products screening in the future. Directing MS/MS spectra acquisition to pre-filtered compounds of interest aids both the classification and identification of unknowns and their prioritization for further characterization as they move along the natural product discovery pipeline. In light of the remarkable progress made over the past couple of years regarding our capabilities for genome mining, instrumental analytics, and computational tools, the result of modern natural products discovery workflows relies more than ever on the tight interplay of all contributing methods and the analytical data used as input for these methods. An almost perfect LC-MS/MS data set can never compensate a genome sequence with bad coverage when aiming for a peptidogenomics approach as it was the case for the new lipothiazole compound class isolated from *Sorangium cellulosum* So ceGT47 (Chapter 4.5.7). Similar problems may be observed when approaching from the opposite direction, i.e. when the genome sequence is exceptionally good but the fragmentation pattern of a compound does not indicate which structural building blocks form the basis of the compound. This is the case for pellasoren where the available genome sequence of So ce38 is almost perfect but the lack of indicative pellasoren fragments prevented the assignment to a gene cluster. Especially compounds related to NRPS/PKS biosynthetic pathways frequently lack characteristic fragments from amino acid building blocks and thereby prevent

their successful classification. As a consequence, unbiased approaches such as fragmentation trees[43] and spectral networking[44] are better choices when it comes to the analysis of highly complex samples without a designated preference for what is searched. These approaches gained momentum in the last couple of years and will certainly evolve to valuable tools for small molecule mass spectrometry.

Final Words

In summary, the newly established method of precursor pre-selection as introduced in this work has the potential to improve advanced MS techniques while the detailed characterization of microsclerodermin and pellasoren biosynthesis contributes to the understanding of microbial secondary metabolite biosynthesis. Although this thesis approached the diverse field of natural product research from two supposedly rather unrelated starting points, it should become clear that both aspects eventually contribute to a screening workflow aiming for the improvement of detection and characterization of new natural products from microbial sources.

5.3 References

(1) Blin, K. et al.: antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013, 41, W204–12, DOI: 10.1093/nar/gkt449

(2) Tang, G.-L., Cheng, Y.-Q. & Shen, B.: Chain initiation in the leinamycin-producing hybrid nonribosomal peptide/polyketide synthetase from Streptomyces atroolivaceus S-140. Discrete, monofunctional adenylation enzyme and peptidyl carrier protein that directly load D-alanine. J. Biol. Chem. 2007, 282, 20273–82, DOI: 10.1074/jbc.M702814200

(3) Li, J. & Jensen, S. E.: Nonribosomal biosynthesis of fusaricidins by Paenibacillus polymyxa PKB1 involves direct activation of a D-amino acid. Chem. Biol. 2008, 15, 118–27, DOI: 10.1016/j.chembiol.2007.12.014

(4) Yonus, H. et al.: Crystal structure of DltA. Implications for the reaction mechanism of non-ribosomal peptide synthetase adenylation domains. J. Biol. Chem. 2008, 283, 32484–91, DOI: 10.1074/jbc.M800557200

(5) Carroll, B. J. et al.: Identification of a set of genes involved in the formation of the substrate for the incorporation of the unusual “glycolate” chain extension unit in ansamitocin biosynthesis. J. Am. Chem. Soc. 2002, 124, 4176–7,

(6) Wu, K., Chung, L., Revill, W. P., Katz, L. & Reeves, C. D.: The FK520 gene cluster of Streptomyces hygroscopicus var. ascomyceticus (ATCC 14891) contains genes for biosynthesis of unusual polyketide extender units. Gene 2000, 251, 81–90,

(7) Watanabe, K., Khosla, C., Stroud, R. M. & Tsai, S.-C.: Crystal Structure of an Acyl-ACP Dehydrogenase from the FK520 Polyketide Biosynthetic Pathway: Insights into Extender Unit Biosynthesis. J. Mol. Biol. 2003, 334, 435–444, DOI: 10.1016/j.jmb.2003.10.021

(8) Wenzel, S. C. et al.: On the biosynthetic origin of methoxymalonyl-acyl carrier protein, the substrate for incorporation of “glycolate” units into ansamitocin and soraphen A. J. Am. Chem. Soc. 2006, 128, 14325–36, DOI: 10.1021/ja064408t

(9) Karki, S. et al.: The methoxymalonyl-acyl carrier protein biosynthesis locus and the nearby gene with the beta-ketoacyl synthase domain are involved in the biosynthesis of galbonolides in Streptomyces galbus, but these loci are separate from the modular polyketide synthase. FEMS Microbiol. Lett. 2010, 310, 69–75, DOI: 10.1111/j.1574-6968.2010.02048.x

(10) Kwan, D. H. et al.: Prediction and manipulation of the stereochemistry of enoylreduction in modular polyketide synthases. Chem. Biol. 2008, 15, 1231–40, DOI: 10.1016/j.chembiol.2008.09.012

(11) Ligon, J. et al.: Characterization of the biosynthetic gene cluster for the antifungal polyketide soraphen A from Sorangium cellulosum So ce26. Gene 2002, 285, 257–267, DOI: 10.1016/S0378-1119(02)00396-7

(12) Zhou, H. et al.: A fungal ketoreductase domain that displays substrate-dependent stereospecificity. Nat. Chem. Biol. 2012, 8, 331–3, DOI: 10.1038/nchembio.912

(13) Häckh, M., Müller, M. & Lüdeke, S.: Substrate-dependent stereospecificity of Tyl-KR1: an isolated polyketide synthase ketoreductase domain from Streptomyces fradiae. Chemistry 2013, 19, 8922–8, DOI: 10.1002/chem.201300554

(14) Keatinge-Clay, A. T.: A tylosin ketoreductase reveals how chirality is determined in polyketides. Chem. Biol. 2007, 14, 898–908, DOI: 10.1016/j.chembiol.2007.07.009

(15) Keatinge-Clay, A.: Crystal structure of the erythromycin polyketide synthase dehydratase. J. Mol. Biol. 2008, 384, 941–53, DOI: 10.1016/j.jmb.2008.09.084

(16) Yadav, G., Gokhale, R. S. & Mohanty, D.: Towards prediction of metabolic products of polyketide synthases: an in silico analysis. PLoS Comput. Biol. 2009, 5, e1000351, DOI: 10.1371/journal.pcbi.1000351

(17) Yates, J. R., Ruse, C. I. & Nakorchevsky, A.: Proteomics by mass spectrometry: approaches, advances, and applications. Annu. Rev. Biomed. Eng. 2009, 11, 49–79, DOI: 10.1146/annurev-bioeng-061008-124934

(18) Scheubert, K., Hufsky, F. & Böcker, S.: Computational mass spectrometry for small molecules. J. Cheminform. 2013, 5, 12, DOI: 10.1186/1758-2946-5-12

(19) Kind, T. & Fiehn, O.: Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics 2006, 7, 234, DOI: 10.1186/1471-2105-7-234

(20) Kind, T. & Fiehn, O.: Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 2007, 8, 105, DOI: 10.1186/1471-2105-8-105

(21) Böcker, S., Letzel, M. C., Lipták, Z. & Pervukhin, A.: SIRIUS: decomposing isotope patterns for metabolite identification. Bioinformatics 2009, 25, 218–24, DOI: 10.1093/bioinformatics/btn603

(22) Pluskal, T., Uehara, T. & Yanagida, M.: Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. Anal. Chem. 2012, 84, 4396–403, DOI: 10.1021/ac3000418

(23) Kind, T. & Fiehn, O.: Advances in structure elucidation of small molecules using mass spectrometry. Bioanal. Rev. 2010, 2, 23–60, DOI: 10.1007/s12566-010-0015-9

(24) Smith, R. M.: Understandin Mass Spectra: A Basic Approach. (John Wiley & Sons, Inc., Hoboken, New Jersey, 2004). 392,

(25) Yang, J. Y. et al.: Molecular networking as a dereplication strategy. J. Nat. Prod. 2013, 76, 1686–99, DOI: 10.1021/np400413s

(26) Rasche, F., Svatos, A., Maddula, R. K., Böttcher, C. & Böcker, S.: Computing fragmentation trees from tandem mass spectrometry data. Anal. Chem. 2011, 83, 1243–51, DOI: 10.1021/ac101825k

(27) Rasche, F. et al.: Identifying the unknowns by aligning fragmentation trees. Anal. Chem. 2012, 84, 3417–26, DOI: 10.1021/ac300304u

(28) Nguyen, D. D. et al.: MS/MS networking guided analysis of molecule and gene cluster families. Proc. Natl. Acad. Sci. U. S. A. 2013, 110, E2611–20, DOI: 10.1073/pnas.1303471110

(29) Watrous, J. et al.: Mass spectral molecular networking of living microbial colonies. Proc. Natl. Acad. Sci. U. S. A. 2012, 109, E1743–52, DOI: 10.1073/pnas.1203689109

(30) Sheldon, M. T., Mistrik, R. & Croley, T. R.: Determination of ion structures in structurally related compounds using precursor ion fingerprinting. J. Am. Soc. Mass Spectrom. 2009, 20, 370–6, DOI: 10.1016/j.jasms.2008.10.017

(31) Heinonen, M., Shen, H., Zamboni, N. & Rousu, J.: Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics 2012, 28, 2333–41, DOI: 10.1093/bioinformatics/bts437

(32) Kersten, R. D. et al.: A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat. Chem. Biol. 2011, 7, 794–802, DOI: 10.1038/nchembio.684

(33) Kersten, R. D. et al.: Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. Proc. Natl. Acad. Sci. U. S. A. 2013, DOI: 10.1073/pnas.1315492110

(34) Ng, J. et al.: Dereplication and de novo sequencing of nonribosomal peptides. Nat. Methods 2009, 6, 596–9, DOI: 10.1038/nmeth.1350

(35) Kavan, D., Kuzma, M., Lemr, K., Schug, K. A. & Havlicek, V.: CYCLONE—A Utility for De Novo Sequencing of Microbial Cyclic Peptides. J. Am. Soc. Mass Spectrom. 2013, 24, 1177–1184, DOI: 10.1007/s13361-013-0652-7

(36) Schmidt, A., Claassen, M. & Aebersold, R.: Directed mass spectrometry: towards hypothesis-driven proteomics. Curr. Opin. Chem. Biol. 2009, 13, 510–7, DOI: 10.1016/j.cbpa.2009.08.016
(37) Lim, H.-K. et al.: A generic method to detect electrophilic intermediates using isotopic pattern triggered data-dependent high-resolution accurate mass spectrometry. Rapid Commun. Mass Spectrom. 2008, 22, 1295–311, DOI: 10.1002/rcm.3504
(38) Yao, M., Ma, L., Humphreys, W. G. & Zhu, M.: Rapid screening and characterization of drug metabolites using a multiple ion monitoring-dependent MS/MS acquisition method on a hybrid triple quadrupole-linear ion trap mass spectrometer. J. Mass Spectrom. 2008, 43, 1364–75, DOI: 10.1002/jms.1412
(39) Rochfort, S. J., Trenerry, V. C., Imsic, M., Panozzo, J. & Jones, R.: Class targeted metabolomics: ESI ion trap screening methods for glucosinolates based on MSn fragmentation. Phytochemistry 2008, 69, 1671–9, DOI: 10.1016/j.phytochem.2008.02.010
(40) Schmidt, A. et al.: An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. Mol. Cell. Proteomics 2008, 7, 2138–50, DOI: 10.1074/mcp.M700498-MCP200
(41) Rudomin, E. L., Carr, S. a & Jaffe, J. D.: Directed sample interrogation utilizing an accurate mass exclusion-based data-dependent acquisition strategy (AMEx). J. Proteome Res. 2009, 8, 3154–60, DOI: 10.1021/pr801017a
(42) Jaffe, J. D. et al.: PEPPeR, a platform for experimental proteomic pattern recognition. Mol. Cell. Proteomics 2006, 5, 1927–41, DOI: 10.1074/mcp.M600222-MCP200
(43) Rasche, F. et al.: Identifying the unknowns by aligning fragmentation trees. Anal. Chem. 2012, 84, 3417–26, DOI: 10.1021/ac300304u
(44) Guthals, A., Watrous, J. D., Dorrestein, P. C. & Bandeira, N.: The spectral networks paradigm in high throughput mass spectrometry. Mol. Biosyst. 2012, 8, 2535–44, DOI: 10.1039/c2mb25085c

Author's effort in the work presented in this thesis

Chapter 2

The author cultivated and extracted So ce38 following the isolation of both pellasoren A and B with full structural characterization. The author hold responsible for the identification, annotation, and full characterization of the gene cluster by analyzing a draft genome of So ce38. All measurements were planned, executed and analyzed by the author (except of the acquisition of NMR spectra). The knock-out mutant was created by Stefan Müller.

Chapter 3

The author isolated microsclerodermin M from a large-scale fermentation of So ce38. Microsclerodermin D and L were extracted from crude extract of MSr9139 which was provided by Suvd Nadmid and Ronald Garcia. All the necessary steps to fully characterize the structure with their stereochemistry such as derivatization, feeding experiments and the analyses of respective results were performed by the author (except of the ozonolysis reaction). The author identified, annotated and characterized the gene cluster in So ce38 and was involved in annotation of the gene cluster MSr9139. The author performed the detailed *in silico* analysis for both gene clusters. All measurements were planned, executed and analyzed by the author (except of the acquisition of NMR spectra). The knock-out mutant was created by Stefan Müller.

Chapter 4

The author planned, measured, and evaluated everything that is part of this chapter except of the bioactivity assays which were performed by Jennifer Herrmann and Viktoria Schmidt.