Nonribosomal peptide synthetases
Zwahlen, Reto Daniel

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
CHAPTER I

Introduction

Reto D. Zwahlen, Roel A.L. Bovenberg,2,3 and Arnold J.M. Driessen1,4

1Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
2Synthetic Biology and Cell Engineering, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
3DSM Biotechnology Centre, Delft, The Netherlands
4Kluyver Centre for Genomics of Industrial Fermentations, Julianalaan 67, 2628BC Delft, The Netherlands

In part to be published as a review in Frontiers in Microbiology, abstract accepted
Abstract

Secondary metabolites represent a class of bioactive natural products and have a long history for human application and are a cornerstone of medicine and chemistry where they have found widespread recognition. Modern medicine has among other factors been enabled through the discovery and therapeutical implementation of secondary metabolites, such as antibiotics or anti-tumor agents and up until this day are a vital component to contemporary medicine. The origin of these products can be traced to plants, fungi and bacteria alike. The underlying biosynthetic clusters frequently contain a multiverse of distinct genes encoding a series of enzymes who interact in a specific manner to ultimately produce a particular product. At the very core of such a cluster is predominantly a complex and multi-modular molecular machinery or more precisely a nonribosomal peptide synthetase (NRPS). NRPS along with related enzymes represent an outstanding class of secondary metabolite producers which have been a focus point in the discovery and engineering of novel bioactive compounds for decades. This chapter will give an introduction to natural products, the associated enzymes and the evolution of accompanying tools for their discovery and engineering with a focus on NRPS.
1. Nonribosomal peptide synthetases (NRPS) and nonribosomal peptides (NRP)

Nonribosonal peptide synthetases (NRPS) are large, highly structured and complex enzymatic machineries, closely related to other modular enzymes such as polyketide synthetases (PKS), NRPS-PKS hybrid synthetase and fatty acid synthetases (FAS). They have certain distinct properties in common, the most striking one being their structural division in domains and modules, which is manifested in their shared evolutionary history (Smith and Sherman 2008). Every enzyme minimally consists of one module, a functionally distinct unit, which allows for the recruitment and subsequent incorporation of a precursor into a growing product. Every NRPS module, initiation (1), elongation (n) or termination (1), requires a minimal set of domains [2]. The two domains essential to every module are the adenylation domain (A) and the non-catalytic thiolation domain (T). This tandem di-domain enables the specific selection and activation of a given substrate. However, the T-domain must first go through 4'-phosphopantetheinyl transferase (PPtase) and co-enzyme A (CoA) dependent activation after expression, by transferring a phosphopantetheine moiety to a conserved serine residue, in order to enter an active state. Also, adenylation domains (A) have accompanying factors, or proteins, called MbtH-like proteins (MLP) [3–4]. In contrast to PPtases, MLPs are merely interacting with the A-domain, however they do not have an intrinsic enzymatic activity, though rather a chaperoning function upon binding a distinct part of the A domain [5–7]. In addition to these domains, any elongation module will require a condensation domain (C), which connects two modules and links up- and downstream activated substrates via a peptide bond. C-domains are stereospecific for both, up- and downstream activated substrates and transfer the resulting intermediate compound to the downstream T-domain. Lastly, the C-terminal termination module essentially requires a thioesterase domain (Te), to catalytically release the covalently bound compound of the NRPS, returning the NRPS complex to the ground state for another reaction cycle. In addition to these essential domains, we can distinguish a series of additional domains, performing epimerization, halogenation, cyclization, macrocyclization, multimerization or methylation [8–10]. Domains as well as modules are clearly defined and evolutionary exchangeable structures amongst multi-modular enzymes. In the case of PKS and NRPS, this lead to the occurrence of a variety of NRPS-PKS hybrids [11–14]. A NRPS can be as simple as a single modular unit containing three domains, although the most complex and largest structure known contains 15 modules with 46 domains [15–16] yielding a 1.8 mD protein complex.
(type I NRPS). Although the size of a NRPS, as well as the modular sequence, limits the size of an NRPS product, it is common that NRPS enzymes cluster and interact with tailoring enzymes in order to produce products of a higher complexity [17]. To enable such specific interactions, NRPS can contain small stretches of up to 30 amino acids at the C- or N-terminus, which form a rather specific recognition point, thus enabling communication (COM-domain) between multiple NRPS of one cluster (type II NRPS) [18–19]. As a result of this multitude of functional dimensions within the NRPS architecture, and the increasing number of NRPS containing biosynthetic gene clusters there is a large growing reservoir of natural products with potentially new properties. NRPS are a crucial part of the secondary metabolism, which is restricted to different prokaryotic, fungal and certain higher eukaryotic species [20].

In comparison to most ribosomally derived peptides, nonribosomal peptides (NRPs) are low molecular weight products. The structural diversity of NRPs is tremendous, mostly due to their chemical complexity. Significantly contributing to this diversity is the fact that NRPS are not only reliant on proteinogenic amino acids only, but up until now more than 500 substrates were identified, which serve as NRPS building blocks [21]. These molecules are predominantly amino acids, but not exclusively, since fatty acids, carboxylic acids and others substrates have been reported in NRPs [20]. NRP thus represent a diverse group of natural compounds and occur as linear, -branched, circular or macrocircular structures [22–23]. The natural functions of NRP are as diverse as their structures. Signaling, communication, metal ion chelation, host protection are important functions occupied by NRP, though many compounds are not fully characterized in this respect. Nevertheless, the characterization of natural products for applied purposes is far developed and led to a vast collection of groundbreaking pharmaceuticals, including antibiotics, anti-fungal agents, immunosuppressants as well as cytostatic drugs [23–25]. The importance of secondary metabolites and hereby also of NRPS, has led to a long history of research aiming at the discovery and understanding of novel NRP compounds and the underlying mechanisms thereof. Natural products and their extracts, including NRP, have been used ever since humanity discovered its beneficial effects. However, only after the discovery and widespread use of penicillins in the 1940s, a rapid expansion in this research field occurred, predominantly focusing on the discovery of novel compounds. After almost four decades, interest in the biochemical and mechanistic understanding of secondary metabolite pathways rose and a ribosome independent way of synthesizing peptides was demonstrated [26]. From then on, the interest in multi modular enzymes and NPRS, respectively, increased exponentially with the availability of novel analytical techniques.
The steady and progressive understanding of NRPS lead to the development of ideas regarding their engineering, potentially allowing for the production of novel compounds with new application spectra. Due to the complexity of the modular assembly line, the utilization of a random mutagenesis approach proved to be unfeasible and its limitations have been shown. However, along with the growing understanding of NRPS and enzymes in general, new ways of directed engineering became available. From a systemic point of view, there are three levels on which engineering may take place, on modules, domains and domain sub-structures or active sites, respectively. Extensive efforts targeting the active site of A-domains has long been considered to be the sole bottleneck in NRPS engineering. Multiple studies confirmed that the substrate specificity of a NRPS A-domain can be successfully altered, however, at the cost of substantially lowered catalytic velocity [27–28]. Similar successes and limitations were observed when natural domains were swapped or replaced by synthetic versions [29]. The most challenging way of obtaining novel NRPS, however, is the swapping or combining of entire modules [30]. Even though this approach seems logical as the module as a functional unit remains untouched, much like the other approaches, it does not take inter-modular interactions across a multi-modular NRPS into account. Directed NRPS engineering remains therefore mostly enigmatic, but in the light of global challenges such as combating drug resistance and sustainable manufacturing, it remains to be a potentially important technology for the discovery and biosynthesis of novel drugs.

2. Evolutionary and functional relationships of NRPS

In spite of the evident relations of the multi-modular enzyme classes of NRPS, PKS and FAS, there are nevertheless functional and structural differences including different places of involvement within the cellular metabolism across these mega-enzymes. While NRPSs and PKSs are exclusively present in the secondary metabolism, FASs are present almost exclusively in the primary metabolism. The differences are not resembled in their respective enzymatic topology, but rather in the corresponding products they synthesize. NRPSs and PKSs can produce compounds containing proteino- and nonproteinogenic amino acids, fatty acids, ketides among other substrates. However, the spectrum of FASs is considerably restricted to fatty acids [31]. This is further underlined by the classes of products formed, for example a majority of antimicrobials are either produced by NRPSs (penicillins, cephalosprins, ramoplanin, enterobactin e.g.) or PKSs (tetracyclines, erythromycin e.g.) [32–33] but none by FAS.
Besides the differences shown above, there is a series of evident similarities between those enzymes, which is best represented in the light of their shared structural features. All of these enzymes can be subdivided into three different types of modules. Overall, product synthesis is universally initiated by activating, starting or loading an initiation module, followed by linear, non-linear or iterative product extension carried out by elongation modules, which may also apply modifications in cis. Ultimately, product finalization and release is catalyzed by a termination module, which contains a release or thioesterase domain, respectively. FAS enzymes can further be divided into two fundamental classes. FAS I a multi-modular enzyme which is present in mammals and fungi, and FAS II, which is abundant in eubacteria and archaea. An alternative classification can be applied to PKSs, which can be divided into at least three distinct classes [33]. All three classes have a similar working mechanism, they all form a product through condensation of building blocks (mostly ketides for PKSs), much like the NRPSs and the FASs. While type I PKSs represent multi-modular enzymes, comparable to NRPSs or type I FASs, type II PKSs are better described as monofunctional multienzyme complexes much like type II FASs. Furthermore, type III PKSs can be distinguished from type I and II PKSs as they lack an acyl carrier protein (ACP) to activate the acyl CoA substrates, thus relying on standalone enzymes for this purpose [33]. Another notable difference in type III PKS is their mostly homodimeric setup, which is rather uncommon among these multi-modular enzymes. The many structural and functional similarities between these complex enzymes suggest that there is a tight evolutionary relationship between them. For example, the phylogenetic relation of the highly conserved ketoacyl synthase domains indicates that FAS I in animals evolved from PKS I of fungi. Those relations are also shown in the overall modular structure, which indicates that FASs and PKSs are, at least structural, more closely related to each other than to NRPSs [14]. Additionally, fungal FAS I presumably evolved through the combination of bacterial type II FAS [34], but all of these enzymes likely evolved from a long process of gene duplication and fusion as well as gene function specification [35]. This universal process of gene fusion and duplication has additionally led to the formation of a diverse class of hybrid synthetases, referred to as NRPS-PKS hybrids [12–13].

The existence of such hybrids further outlines the evolutionary picture of the close NRPS-PKS relationship, leading to functional enzymes which are highly active during the bacterial and fungal secondary metabolism. Further details elucidating structural and functional features of NRPS will be described in the respective sections of the subsequent paragraph.
3. NRPS domain function and biochemistry

3.1. Adenylation and thiolation domains

Any NRPS module minimally consists of an A- and T-domain, enabling single module functionality and multi-modular functionality upon addition of C domains [8;36–37]. A-domains are roughly 500–550 amino acid in size and in general determine the substrate specificity of a module due to their specific substrate recruitment [38–41]. They are often referred to as “gatekeeper” domains, as there is no subsequent product formation without prior adenylation and thioesterification of a substrate [42] (Figure 1). Bioinformatic prediction of the precise specificity of an A-domain is a delicate and still error prone process. However, a series of 10 core motifs (A1–A10), essential to all NRPS A-domains have been elucidated and in part functionally assigned [43–44]. Based on this sequence motif, a core or “Stachelhaus” code has been determined, which covers 10 positions mainly across the A4 and A5 motifs and describes the essential residues that determine the adenylation specificity [39]. Most of the identified motifs, core motifs and Stachelhaus code, both predominantly cluster in and around the active site of the A-domain, but not exclusively [45–6]. Furthermore, any A-domain can be further subdivided into two distinct parts, the Acore and Asub part [47]. This division is mainly across catalytically inactive parts, however, due to its dynamics, still strongly affects the active site specificity [48]. In addition to the 10 core motifs, a 11th has been proposed, located at the junction between the Acore and Asub parts, seemingly fulfilling an essential hinge function, manifested in the conserved LPxP motif [49] Structural flexibility is crucial in the sequential tandem reaction of the A-domain. The two core functions of the A-domain are characterized first, through the hydrolysis of ATP or adenylation, allowing an AMP-substrate conjugate to be formed, which is subsequently transferred to the free thiol group of the 4′-phosphopantetheinyl-moiety (ppant), which is anchored to a conserved serine residue in the downstream T-domain [50–52]. This thioesterification is guided by a structural rearrangement of the Asub-domain, which itself undergoes a rotation of about 140° [53–54] relative to the Acore-domain. This process is dynamically well conserved and can also be observed in other enzymes of the adenylate-forming enzyme superfamily (ANL), such as acyl-CoA synthetases and luciferases [55], where this mechanism has been primarily characterized [53;56]. However, not only the dynamics, but also the structure among this family of enzymes appears conserved in a relatively strict manner. Especially the small Asub domain seems to bear an extremely important function, not only
Figure 1 — Adenylation and thiolation domain functionality.

(A) Sfp chaperones the transfer of the phosphopantetheine moiety from CoA onto a conserved serine residue in the thiolation domain (GGxS). The NRPS in the activated holo-state subsequently recruits free amino acids through the adenylation domain (B), which forms an amino acid — AMP adenylate intermediate, ultimately leading to the transfer of the adenylated amino acid onto the ppant-arm in the thioesterification conformation.
in chaperoning the transition between the adenylation and thioesterification sub-steps within the A domain, but also in respect to the displacement of the T domain [48;53;57]. The 4 α-helix containing T domain needs to partially penetrate the A domains active site in order to link the activated substrate to the carrier ppant arm and subsequently donate it to the downstream domain. This process is indvertibly linked to the precise movement and action of the upstream A\textsubscript{sub} domain.

### 3.2. Condensation and epimerization domains

Condensation (C) domains are approximately 450 residue NRPS domains, representing a highly versatile class of NRPS domains. Any NRPS composed of more than one module must consequently contain at least one C-domain. However, also single modular NRPS may contain C-domains, especially if they cooperate with other NRPS. Essentially, the primary target of a C-domain is the condensation of the up- and downstream activated substrates through a nucleophilic attack, mainly leading to the formation of a di-peptide linked via a peptide bond (Figure 2). Nonetheless, several residues of the C-domain may have the intrinsic potential to fulfill multiple functions. The biochemical spectra of these pluripotent domains includes double functions of condensation and epimerization, N-methylation, cyclization or serine-dehydration [58–60]. C domains are classified among the chloramphenicol acetyltransferases (CAT), sharing highly conserved CAT-like folds in their structure. Nonetheless, as opposed to the commonly observed trimer formation of CAT enzymes, C- and C-like domains are monomeric and resemble a pseudo-dimeric V-like structure harboring the active site at its center [61]. In order to gain access to the up- and downstream activated substrates, the active site is accessible through two tunnels, which ultimately connect at the active site in the vicinity of the essential C-domain motif HH\textsubscript{xxx}DG [20]. The second histidine of this motif is essential for the peptide bond formation, as it appears to be involved in the orientation of metabolic intermediates, thus preventing the dissociation of the substrates from the active site. This function can also be seen as a proofreading step, which eliminates incorrect intermediates from the assembly line.

Epimerization or E domains are among the most abundant modification domains intrinsic to NRPS. Interestingly, they are structurally closely related to C domains which is reflected in their pseudo-dimeric structure as well as their active site position and composition [61–62]. In contrast to C domains however they are responsible for the site specific epimerization of a
Figure 2 — Linear assembly of a tripeptide on a NRPS.

(A) Following the successful transfer of all required amino acids onto their respective thiolation domains, they are handed to the up- or downstream condensation domain, starting with the N-terminal module, in linear acting NRPS. (B) The resulting di-peptide remains covalently attached to the downstream T-domain, allowing for the poly-peptide intermediates to be extended with additional amino acids. (C) The final product is ultimately released through a thioesterase domain, laying on the C-terminus of the NRPS.
substrate, predominantly carrying out this function after peptide binding has occurred. Furthermore, the active site containing a central histidine among other bulky residues, allows it to fulfill a similar proofreading function in the metabolite production as seen in C domains, as it may be aiding in the process eliminating incorrect intermediates [8]. This process is crucial in the production of active compounds, since even minor alteration to the stereochemistry of a compound can render it completely inactive. Lastly, despite only few available structures, biochemical studies [36;61–65] have shown clearly that these domains exhibit considerable specificity and are in a fine balance with the adjacent domains, which is further underlined with double C-E functional domains.

3.3. Thioesterase and other domains

In addition to A-, T- and C-domains, which are essential to any multi-modular assembly line, there are additional domains enabling on-line product modification and most importantly product release, thus allowing for another catalytic cycle. Ultimately, alterations applied by intrinsic NRPS domains, as well as product release related structural modifications lead to either partial or full functionalization of NRPs or allow for product recruitment through intrans acting tailoring enzymes.

3.3.1. Product release and functionalization

Thiotemplate based enzymatic systems rely on a catalytic activity in order to remove a product or product-scaffold of the primary enzyme. Therefore, most NRPS contain a domain on their C-terminus responsible for precisely this purpose, the thioesterase domain (Te). Te-domains are a common commodity in single and multi-modular NRPS, although, in multi-NRPS systems only the terminal NRPS carry this domain. For some assembly lines, type II thioesterases fulfill this function representing a standalone alternative to intrinsic domains. Fundamentally, three different release mechanisms can be distinguished, linear product hydrolysis, intra- or inter-molecular cyclization or substrate multimerization. All reactions underlie a series of sequential steps in order to allow for proper compound release, sharing an intermittent step in which the product is covalently linked to a conserved serine residue in the Te-domain ser-his-asp motif [9]. This mechanism is similar to T-domain linkage, though in contrast to the T-domains, Te-domains bear an intrinsic
catalytic activity. This represents a special case among NRPS domain and is highly related to serine hydrolase mechanisms. Structural data suggests that product transfer from the T-domain onto the serine of the Te-domain leaves a significant amount of free space within the active site tunnel which in turn allows for complex compound rearrangements. Basic linear hydrolysis appears to be the most unspecific form of product release but it is also prone to unspecific reactions, such as the release of unfinished product intermediates or single substrates from the upstream T-domain. Nonetheless, it represents an efficient way of product release and can be observed in the production of the penicillin precursor ACV [66] or feglymycin [67–68]. Cyclization reactions on the other hand seem to be carried out in a more directed and specific way. The higher reliability of these reactions can mainly be attributed to the increased coordinative need in order to accommodate a cyclization of a long, charged and/or bulky substrate. An evident case, where this proofreading or gate-keeping is performed by the Te-domain is the β-lactam antibiotic nocar dicin. The product of the two NRPS NocA and NocB, undergoes monocyclic β-lactam ring closure, where the Te-domain serves as a gatekeeper, holding the product until ring closure has been finalized [69–71]. Similarly, stringent control mechanisms can be seen in macrocyclization as well as in multimerization reactions, as for example with the production of gramicidin S. Overall, the perspective on Te functionality has been gradually developed, revealing a both dynamic and specific domain, which serves as a catalyst to a wide range of highly time dependent product functionalization and release reactions [9;70].

3.3.2. **Intrinsic product modifying domains**

The focus of domain identification and characterization has long focused on the previously established and described domains with in-detail studies on their linked mechanisms. However, recent studies have not resolved complex sequential interactions and catalytic cascades, but have also deepened our understanding of the mechanistic roles of the underlying domains as well as their characteristics, respectively. In addition to the C-domain related Epimerization domain, discussed previously, there are cyclization (Cy), oxygenation (Oxy) as well as methyl-transferase (MT). These domains have been characterized to the extent of classifying their functions, although, especially Cy- and Oxy-domains may occur as a singular bi-functional unit or in a serial manner, respectively [72]. The aforementioned domains are integral parts of a NRPS assembly line, in contrast to recruited domains and
enzymes, which are discussed at a later point. Their intrinsic nature indicates not only that the reactions performed are of utmost importance with respect to product functionality, but also points towards a very archaic and conserved function of the domains. Cy- and Oxy-domains in fact, specifically replace the classic function of C-domains, omitting amino acid condensation through peptide bond formation, resulting in thiazoline, oxazoline or methylloxazoline structures [72–73]. Those reactions predominantly occur in siderophore producing NRPS and rely on the presence of serine, threonine and cysteine residues [74–75]. Recently, a more detailed, structure based comparison of a Cy- with C-domains delivered further support on the common ancestral history of the two domain types, shown in the overall conserved scaffold features and accentuated differences clustered across the active site of the domains, respectively [76]. Also, MT-domains follow the common di-sub-domain structural patterning, which is also seen in A-, C-, E- and Cy-domains. Fundamentally, MT-domains, however, are more restricted in their functional spectra, which covers the transfer of methyl-groups from S-adenosylmethionine to N (N-MT), C (C-MT), O (O-MT) or for certain residues S (S-MT) atoms resolving around the amino acids Cα carbon [77] and in case of S-MT Cβ, respectively [78]. In NRPS the most abundant MT specificity observed and described is N-methylation, although an increasing number of domains have been proposed which carry out C-methylations. No conclusive O-methylation activity has been shown up to date [10].

4. NRPS multi-modular structures and dynamics

In contrast to the biochemical characterization of NRPS domains as well as entire assembly lines, there is considerably little structural insight available. Even though all types of domains have been structurally characterized by means of X-ray crystallography or NMR, the elucidation of distinct conformational states of a given domain remains a challenging task. This is especially a limiting factor with respect to the few multi-domain structures which have become available, as it limits the implications that can be derived from such novel structures. Furthermore, due to the countless conformational and dynamical changes NRPS modules and domains undergo, it has been impossible so far to retrieve any structure of an entire multi-modular NRPS assembly line. Nonetheless, on basis of the spatial knowledge obtained in the process of module and domain characterization, efforts have been made to describe the catalytic sequence, timing and dedicated dynamics.
4. NRPS multi-modular structures and dynamics

4.1. NRP synthesis and the associated conformational implications

Any given NRPS linked product synthesis begins with the adenylation of incoming amino acids, aided by an A-domain in its initial adenylate forming conformation. The subsequent rotation of the C-terminal $A_{\text{sub}}$-domain, allowing to adopt the thioester forming conformation appears not only to be the limiting step in product formation [53;55], but furthermore guides the ppant arm into the active site tunnel of the downstream C-domain [79]. This finely timed and interlinking conformational changes are further manifested in the domain architecture, as there are dedicated non-conserved areas in the T-domain, which appear to bear dedicated functional relationships on a structural level with the A- and C-domains up- and downstream.

Figure 3 — Potential structural arrangement of a tri-modular NRPS.
Shown in A is the potential side view of a three modular NRPS on top and an EM image on the bottom. B displays the top view of the same NRPS along with the view of the particle through EM on the bottom. Both schemes are adapted from [81].
within the assembly line, respectively [62]. The clockwork like intertwining domain-domain interactions continue beyond the borders of a given module as they appear to extend into the downstream modules and their C-domain. This is proposed, as C-domains, representing the first or N-terminal domain of a module, appears to be depending on the movement of the upstream T-domain in order to rearrange into conformation and proximity, ultimately allowing for the formation of a peptide bond [48;80]. The fundamental dynamics of how a polypeptide is synthesized within the framework of a NRPS seems to be a common although also highly case dependent mechanism, but ultimately should lead to a transferable concept to all NRPS. However, taking product release and functionalization (as discussed in part III) as well as the multitude of additional, internal and external modifying domains into this greater dynamic scheme, the situation becomes infinitely more complex as a result of the great number of different NRPS architectures and involved factors. The overall interactions can only be conclusively determined with novel, multi-modular structures of NRPS, which are non-existent as of this moment. Nonetheless, there have been attempts at creating multi-modular models, based upon the crystal structures of single NRPS domains and the interfaces known through the solution of whole module structures. Most notably the model proposed by Mahariel [81] has shed light on this issue and underlines the high level of organization within multi-modular NRPS, leading to the standing theory of a helix like domain setup, resolving around centrally aligned T domains (Figure 3).

5. NRP or NRPS associated factors

In order to create the diversity and complexity of compounds observed in NRPS, a series of accompanying factors are required in addition to the intrinsic NRPS domains. Two main actions can be distinguished within these factors, NRPS chaperoning or activation and compound targeting elements. The first category includes enzymes which are either strictly essential such as the phosphophantetheiyl transferase sfp or may enable or improve catalytic properties as it is the case for the MbtH-like proteins (MLP). The latter group of factors ultimately target the growing NRP and may act during catalysis, such as factors which alter a compound during the NRPS associated state or once the compound is released from the assembly line. The latter concern the tailoring enzymes, as they vastly contribute to the very diverse structures observed in NRP.
5.1. Tailoring enzymes

Secondary metabolites exhibit diverse scaffolds, shown as linear, side-branched, circular or macro-circular structures which is determined by the ability of the NRPS to generate these compounds in a linear or iterative process. There is a series of site directed alterations which have been evolutionary fine-tuned to enable more potent and specific properties. Certain residue alterations and rearrangements can be attributed to intrinsic NRPS domains (see section III), though many functions are absent of the NRPS assembly lines and require external tailoring enzymes. Genes encoding tailoring enzymes may be clustered among the NRPS gene clusters, though many are derived from other pathways [82]. Two main interactive product alteration mechanisms can be distinguished, NRPS-associated and NRPS independent, or on-line and off-line, respectively. On-line modifications target either substrates or compound intermediates, which are in a covalently-linked state with the host NRPS assembly line and can be essential for product formation [73]. This type of modification is either carried out by an intrinsic domain or more relevant in this section, a standalone tailoring enzyme or domain. Recently, the recruitment of cytochrome P450 enzymes has gained significant attention [60;83–85]. Genes of oxygenases and especially P450 enzymes frequently cluster among NRPS genes and the resultant enzymes have been shown to carry out substrate modifications, for example in skylamycin [86] where a single oxygenase performed three hydroxylation reactions [87]. More interestingly, NRPS association and substrate modification appears to be highly selective and it has been shown that the T-domains play a crucial role in P450 recruitment [83;88]. Another series of experiments targeting teicoplanin, a glycopeptide antibiotic, revealed that not only T-domain targeting is used for P450 selection, but also a novel type of domain, the X-domain, which appears to be closely related to C-domains, although with an active site rendered unusable for peptide bond formation. The example is even more outstanding as this reaction requires four dedicated oxygenases [60].

Another modification essential for product activity is the halogenation of residues. Halogenation may be carried out on-line or off-line [89], but always requires a standalone halogenase. The most abundant form of halogenation appears to be the chlorination of benzoic ring structures as it is seen with teicoplanin, enduracidin or ramoplanin [17;90–91]. Moreover, there is a series of modifications, namely glycosylations, acylation and sulfation, which are generally restricted to off-line mechanisms. Glycosylations are highly abundant for many NRPS products, especially glycopeptides. They are often not essential to the biological activity of the molecule, though they may improve
solubility as well as stability of a compound [92]. In contrast, neither acylation nor sulfation appear to be a widespread phenomenon. Acylation is attributed to a new class of N-acyltransferases, which were only recently classified and structurally analyzed [93]. Sulfation reactions are equally rare and associated sulfotransferases are rather uncommon, however the records have been expanded steadily in the past few years [82].

5.2. MbtH-like proteins (MLP)

Another potentially essential factor in the line of NRPS compound production is the MbtH-like protein (MLP). The MLP is currently the only catalytically inactive interaction partner known for NRPS, and they have been rather systematically analyzed. MLP proteins are ~70 amino acid peptides arranged in three β-sheets, and in most cases one N-terminal α-helix [94–95]. Through association with the A-domain of an NRPS the MLP appears to increase the kinetic properties of the interacting A-domain. In addition, they may even improve NRPS expression and translation levels and rates, respectively, using a non-characterized mechanism [5]. Potential interaction sites have been characterized [6;97]. It appears that the MLP associates in a dynamic way with the rigid Acore-domain, and to a lesser extent via weak intermolecular interactions with the Asub-domain. Available structures clearly show that three MLP borne conserved tryptophan residues align with two alanine residues on the Acore-domain. Based on a series of studies, investigating the transferability of MLP homologues between different prokaryotic organisms and strains, a minimal MLP functionality motif was created [95;98]. Although structures have been determined and important residues have been identified, there appears to be no report of MLP presence in eukaryotic organisms, despite the widespread occurrence of NRPS across most sequenced prokaryotic organisms [3].

5.3. Phosphopanthetheinyl-transferases (PPTases)

PPTases form a distinct enzymatic family and are essential to the functionalization of NRPS, FAS as well as PKSs. Fundamentally, they covalently link the CoA derived phosphopanthetheine moiety onto Carrier proteins of the aforementioned megasynthetases. They target NRPS thiolation domains or FAS and PKS acyl carrier proteins (ACP) with various specificities. Overall PPTases can be divided into three distinct sub-types, (I) ACP specific holo-ACP
synthase (AcpS) targeting FAS and PKS, (II) Surfactin-phosphopantetheinyl transferase (sfp)-type NRPS specific PPTases and (III) the FAS-fused domain like megasynthase PPTases [99]. All three classes appear structurally diverse. Type I PPTases form a homotrimer of about 360 aa containing three active sites at the interfaces of the trimer. In contrast, type II structures are monomeric 240 aa enzymes, although containing two equally sized domains, which resemble a pseudo-dimer containing a single active site. Based on the available crystal structures as well as primary sequence analysis, it is likely that type II PPTases arose through gene duplication and fusion from the archetypic type I PPTases. Type III PPTases however, appear to have very unique structural and multimeric states, resembling significantly more complex structures of homo-hexamers as well as hetero-dodecamers in a size range of 0.54–2.6 mD [100–101].

For NRPS related work however, only type II PPTases are of importance, as they are the only one which exhibit a distinct specificity towards T domains, while retaining sufficient promiscuity to target different NRPS T domains as well as CoA similar analogues, respectively. Type II PPTases are ubiquitously found across prokaryotic as well as eukaryotic species and often cluster among NRPS genes, although one copy is predominantly sufficient for one organism. Secondary metabolism is heavily impaired upon deletion of type II PPTases and purification experiments under PPTase non-overexpressing conditions showed extremely low PPTase and NRPS expression levels, further hinting at the essential function, high affinity and activity of this PPTase type. Overall, PPTases require a significant amount of experimental attention, as neither conformational dynamics, specificity prediction nor the multimeric state are clarified at this point, though the utilization of broad spectra type II PPTases such as sfp are sufficient to bypass major bottlenecks in the design and engineering of NRPS based pathways [102].

6. Analytical tools and methods for NRPS and NRP identification and characterization

6.1. Metabolite and product discovery

The most direct route to discover NRP is a functional screening, as it has been applied ever since the early 1930s for the discovery of first generation antibiotics and the subsequent screening of improved producer strains [103–105]. Exploiting the intrinsic properties of a product however is equally elegant, as it is limited to targets which have a considerably high potency. This has been
established for antimicrobials, where sensitive indicator strains are used in combination with halo based plate or broth assays for example [106–109]. A similar concept can be applied to the discovery of pigments such as indigoidine, as changes in the color of a media can be readily determined with spectrophotometrical devices [110–114]. In addition to the mere discovery, pigment producing NRPS such as IndC of the indigoidine pathway, have been successfully employed as biosensors for combinatorial engineering efforts [29]. Furthermore, associated strategies relying on the ability of the host to produce scavenger molecules such as siderophores enable the utilization of growth properties as a phenotype to evaluate suitable targets. In addition to the intrinsic properties a series of combined reactions, using indicator substrates or secondary catalytic reactions significantly contribute to the broad range of bioassay applications. This type of assay may also make use of binding proteins as shown for the penicilllin binding protein or analogous mechanisms [115–117]. However, despite the potential for high throughput and quantitative screening, this method is merely applicable to a minority of desired compounds, often because of their low abundance or the lack of a directly or indirectly observable effect. This presents an even more eminent problem with anti-cancer drugs, as a tremendous selection of indicator strains must be evaluated in order to near conclusively determine their potential.

Another very direct and versatile technique to determine the presence of secondary metabolites is the analysis of supernatant or other types of samples containing excreted metabolites is mass spectroscopy, coupled with liquid chromatographic techniques (LC/MS). Even though the technique per se is not novel, recent developments were aiming at methods which allow for a more sensitive and overall inclusive detection. An additional advantage is that not only fully functionalized products, but also side-products, and metabolic intermediates can be analyzed and characterized. Though the increased depth of analysis also increases the time investment, limiting it to low-medium throughput screening. However, with ever decreasing costs for analytics and the parallel development of more reliable and sensitive techniques, this approach poses a serious alternative especially for large scale users.

Nonetheless, all these methods may be useless for very low abundant metabolites as well as any silent biosynthetic gene cluster. Though these metabolites potentially account for a significant amount of total natural products and should not be neglected. Luckily there are several ways for the improved or induced production of low abundant or cryptic metabolites, respectively. Based on the identification of promising biosynthetic gene clusters, discussed in the following paragraph, approaches targeting promoter sequences or accompanying factors such as PPTases or MLPs may aid in
increasing or triggering the production of novel metabolites. Lastly, in case of partially understood pathways, the simple feed of low abundant substrates can improve the yield of a given pathway product.

### 6.2. Bioinformatic and site-directed tools

The discovery of conserved and highly specific motifs as part of many functionally specific NRPS and NRPS domains has laid the fundament for the development of specific discovery tools. NRPS are frequently encoded in large genes with very dedicated genomic features. The combination with continuously more affordable genome sequencing techniques, allows for the relatively simple identification of putative biosynthetic gene clusters, based on their DNA sequence [118–123]. Due to their modularity and dedicated functional domains, NRPS can be identified through homology based searches, which allow for the identification of NRPS by means of domain sequence and more precisely using the conserved active site motifs for A (A1–A10 core motifs), T (GGxS), C (HHxxDG) domains [124–126]. Furthermore, product modifying domains may be identified in an analogue manner, allowing to predict the full functionality of a NRPS based upon the primary DNA sequence. This prediction can be further taken into account for the prediction of NRPS specificities and thus also allow for the determination of hypothetical products and product intermediates. Other NRPS-associated hallmarks, such as MLP, PPTases or tailoring enzymes contribute to a more global cluster identification. In fact, the latter factors, especially type II sfp-like PPTases, are a strong global indicator of functional secondary metabolite clusters, giving indication of secondary metabolite production by an organism.

Even a step further, towards the discovery of specific parts and their partial isolation is the development of domain specific DNA/peptide signature sequences, allowing to identify NRPS parts at proteomic level [127–128]. Such probes generally consist of two parts, one part which associates with a specific domains active site and a second part which allows for the detection of the domain-probe conjugate [129–130]. The domain specific parts may target an amino acid or an adenylated substrate for A and T domains respectively, including non-natural substrates. The indicator entity of this probe on the other side may be an affinity tag, i.e. biotin, 6×his, a fluorescent probe or a radiolabeled group [129;131]. A combinatorial approach using these probes along with electrophoretic methods, such as capillary electrophoresis or 2-D gels, represents a powerful tool to obtain direct proof of an organisms potential to express a specific NRPS. Lastly, this approach additionally allows...
for the extraction of specific NRPS domains, which may subsequently be used for dedicated engineering efforts.

7. Engineering of NRPS: targets, strategies and challenges

NRPS are highly structured and multi-facetted enzymes, bearing a tremendous potential in the exploitation of their scaffold for the generation of novel, bioactive compounds. However, due to the complexity of all interactions within these mega enzymes, the elucidation and implementation of engineering strategies is an extremely challenging task. Several strategies have been developed and applied with different degrees of success, though the overall approaches can be grouped as module, domain, sub-domain or site directed, respectively. All of these strategies have their inherent difficulties, advantages and disadvantages in respect to the complexity and success rate of NRPS engineering efforts and will next be highlighted with selected examples.

7.1. Subunit, module and domain swapping

Due to the strict arrangement of NRPS in domains and modules, the possibility of exchanging a sub unit appears to be the most straight forward approach for altering the intrinsic properties of a NRPS. A series of studies targeting the enzymes linked to the production of daptomycin [132–133] elucidated the possibilities and borders of a combinatorial swapping strategy in light of novel compound production. The daptomycin biosynthetic cluster comprises of three NRPS containing a total of 13 modules for the incorporation of an equal number of substrates. Different levels of domain and module swap approaches were followed, starting with the exchange of modules 8 and 11 (C-A-T), representing an internal module exchange. The resulting NRPS exhibited the production of novel daptomycin compounds with an inverted amino acid composition at the predicted sites at a near native rate. However, if the less homologous epimerization domain (C-A-T-E) is taken into this swapping approach, the production levels drop under 1 % of the original activity, further illustrating the importance of precise inter-modular interactions [132]. Subsequent experiments targeting the third daptomycin NRPS DptD, focused on the replacement of the entire protein with highly homologous units of the calcium dependent antibiotic (CDA) and A54145 pathways, respectively. Despite significant production levels of novel compounds, no increased anti-microbial spectra nor activity was unraveled. It
appears that all hypothetical lipopeptides which may result from the given NRPS have been through extensive evolutionary selection, thus leading to daptomycin as the most suitable compound for the host organism. A similar combinatorial approach has been chosen for the investigation of pyoverdin and potential analogues, resulting in more moderate production rates of analogue compounds, likely due to a more disruptive engineering strategy in respect to the conserved inter-modular interactions and interfaces [134–135]. A different approach on the improvement of a NRPS activity was taken with the indigoidine syntehase $IndC$. In contrast to the majority of studies targeting single domains for exchange purposes, not the A- or C-domain was chosen, but rather the significantly smaller T domain [29]. In a remarkable effort, not only functionally comparable NRPS variants with homologous T-domains from other organisms were created, but additionally, synthetic T-domains were successfully utilized to create novel NRPS which surpass the native kinetic potential of $IndC$. Overall, the choice of a particular approach must be tailored to the availability of homologous parts and if applicable, a series of domains, or an entire module should remain intact in order to circumvent the alteration of crucial inter-domain and module interactions. Nonetheless, combined computational and swapping approaches, may bear a significant potential for the improvement of NRPS, but are still heavily limited by the lack of structural and experimental data.

### 7.2. Sub-domain exchange and directed evolution

Next to engineering efforts to the domain and module borders, there is a series of approaches which directs attention to a range of sub-domain parts, including active sites, binding sites as well as even more minimal changes. While even partially directed approaches still require the generation of enormous mutant libraries, a recent study has drawn significant attention towards a more elaborate exchange strategy, targeting binding pocket related regions [30;136]. In essence this approach is based on an event which has occurred at least once in the NRPS evolutionary history [137]. Based on this observation, a region was targeted in the gramicidin S synthetase A $GrsA$, which spans across most relevant binding pocket residues. The exchanged parts from either the NRPS $GrsB$, whose gene is located downstream from $grsA$, or other microbial sources, resulted in a series of active chimeric NRPS. Moreover, neither specificity nor catalytic activity were dramatically reduced, thus delivering a potent scaffold for further improvement using a computational and directed evolutionary strategies.
7.3. Engineering challenges, application potential and outlook

An increasing level of detail about the structure and functions of the modules and domains as well as about biosynthetic capabilities of NRPS are being unraveled. Hereby, both the diversity of the NRPS as well as their potential for biotechnological applications becomes apparent, resulting in an increasing number of products from the secondary metabolism of bacteria and fungi being considered or already used for societal purposes. Due to the massive genome sequencing efforts, also the overall number of secondary metabolites, biosynthetic clusters as well as potential producing organisms is increasing continuously. Nonetheless, facing global health challenges, such as the anti-microbial resistance, anti-cancer and other drugs, require not only the discovery of new NRPS derived products and characterization of organisms, cluster and products, but also the development of enabling techniques for the exploitation of the full biochemical potential of their dedicated NRPS. Even though there is a general idea concerning the structure of NRPS modules, domains and their biochemical functions and specificities, NRPS engineering efforts have been limited, as discussed previously. Nonetheless, the potential biochemical diversity provided by this enzyme class is tremendous and ideas how to unlock this potential have been proposed and will be further discussed.

To provide for a fundament to an interlinked and systematic platform to explore the full potential of NRPS, two bottlenecks must be considered. The first concerns the ever growing reservoir of potential targets, gene clusters and standalone NRPS, which have been fully or at least partially identified and characterized. The second concerns the partial and incomplete structures of NRPS domains and modules which have shed some light on essential sub-structures, without revealing overall structures and dynamics of multi-modular NRPS systems. Due to the complexity of this class of mega-enzymes, a mere random engineering approach seems disproportionally laborious and will eventually require a target and site preselection within a given NRPS assembly line, based on a combinatorial approach. Such an approach should allow for a preliminary target selection, based on the proposed function of the projected novel enzyme and a sequential selection, allowing for the proposition of sub-structural targets for a more directed engineering approach.

A platform as described above should furthermore be accessible through various inputs, covering the entire range from primary genetic code, discovered in metagenomics samples, till metabolics data and bioinformatic predictions of novel hypothetical structures of a proposed compound found in
fermentation samples. AntiSMASH software and database [118] illustrates how the fundament of such a platform may be envisioned. Many features, including the predicted architecture of NRPS domain topology, the assignment of tailoring enzymes and a crude proposition of the associated metabolic product are indeed already included. However, for example promoter, terminator and RBS properties or domain specificities are yet to be determined using a manual approach. While the latter expression features may be implemented in such a platform with relative ease, by the addition of external links or the embedding of additional software tools and databases, an entire second software layer, focusing on structural features or predictions of the associated enzymes is missing. Naturally, due to the limited amount of structural data available, this feature would heavily rely on homology based structural models. Nonetheless, the systematic nature of the platform has the potential to increase the speed and precision of the given predictions while simultaneously expanding the internal database vastly. In an ideal setup, this process may even allow for the discovery of more universally transferable NRPS domains or generally structural elements, which could ultimately be used in a transferable parts library, similar to the BioBricks concept [138–139]. The potential success of such a platform however, will eventually rely on the overall ease of structural data acquisition and more specifically on the possibility to solve any multi-modular NRPS structure.

8. Scope of the thesis

The overall aim of this work concerns the fundamental characterization of the NRPS ACVS and the application of these insights in engineering a hybrid NRPS to synthesize hpgCV. This hybrid hpgCV NRPS would enable the development of a new biosynthetic pathway, for the production of the antibiotic D-amoxicillin in a fully fermentative manner. In order to establish this novel NRPS activity, an A-domain is required capable of recognizing hydroxyphenylglycine (hpg). Chapter II describes the identification and characterization of a multitude of hpg activating domains and demonstrates a system, which allows for the screening of such domains in a medium-throughput fashion, bypassing the utilization of radiolabeled substrates. Using the previously characterized domains, a series of chimeric NRPS were constructed. Therefore, a versatile cloning and expression system was established allowing for the dedicated and convenient exchange of NRPS domains in a larger enzyme (Chapter III). The ACV synthetase of Nocardia lactamdurans served as a template for this system due to its natively beneficial specificity and
compatibility with heterologous expression in *E. coli*. For a more complete understanding of the ACVS a series of biochemical methods were developed and applied. In addition, structural features of this three modular ACVS NRPS were assessed providing novel data on the overall structural dynamics (Chapter IV). A selection of chimeric hpgCV hybrids were subsequently transferred to the production host *Penicillium chrysogenum* and together with NRPS associated protein MLP examined for the production of D-amoxicillin and the tripeptidic precursor hpgCV (Chapter V). Finally, the effects of different MLP of prokaryotic sources were evaluated with respect to a series of *P. chrysogenum* native NRP products (Chapter VI). In the outlook, new NRPS engineering strategies are discussed.

9. References


64. Zane HK, Naka H, Rosconi F, Sandy M, Haygood MG, Butler A. Biosynthesis of amphi-enterobactin siderophores by Vibrio harveyi BAA-1116: identification


76. Bloudoff K, Fage CD, Marahiel MA, Schmeing TM. Structural and mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis. Proc Natl Acad Sci U S A. United States; 2016;113: 1614191114


100. Beld J, Sonnenschein EC, Vickery CR, Noel JP, Burkart MD. The phosphopantetheinyl

References


127. Chen Y, McClure RA, Kelleher NL. Screening for expressed nonribosomal peptide synthetases and polyketide synthas-


