Universiteit Leiden
The Netherlands

Expanding the chemical space of antibiotics produced by Paenibacillus and Streptomyces

Machushynets, N.V.

Chapter 2

General introduction

Pioneering and contemporary challenges in antibiotic discovery

The discovery and introduction of antibiotics into clinical use was the most remarkable medical breakthrough of the past century. Fleming's discovery of the first antibiotic, penicillin, in 1928 marked the beginning of the modern medical era of antibiotics (Katz & Baltz, 2016, Fleming, 2001). Inspired by Fleming's discovery, Selman Waksman started a systematic study of microbes as producers of antimicrobial compounds in the late 1930s (Lewis, 2012). The canonical "Waksman platform" for antibiotic discovery involved the identification of new bacterial isolates from the soil samples, antimicrobial screenings, and subsequent bioassay-guided purification of the bioactive compound (Waksman, 1945). Waksman discovered fifteen antibiotics made by soil-dwelling actinomycetes, including streptomycin, the first antibiotic against tuberculosis. Selman Waksman's pioneering work sparked a significant turning point in antibiotic discovery and opened the door to the golden era of antibiotic discovery. This period witnessed a surge in the identification and development of crucial antibiotics like bacitracin, tetracycline, polymyxin, vancomycin, and many others (Katz & Baltz, 2016) (Figure 1). More than half of all the classes of clinically used antibiotics came from actinomycetes, while the rest derived from fungi and Firmicutes (Katz & Baltz, 2016).

Soil bacteria produce a wide range of natural products that protect against insects, herbivores, and phytopathogens (including bacteria, fungi, nematodes, and viruses), as extensively highlighted in numerous reviews (Katz & Baltz, 2016, Bérdy, 2005, Newman & Cragg, 2020, Olishevska *et al.*, 2019). Several theories have been proposed to explain why soil microbes produce such a diverse array of bioactive natural products (NPs). The most likely explanation is that they have multiple functions and act as chemical weapons or signaling molecules; or mediate interactions with eukaryotic hosts such as insects and plants (Seipke *et al.*, 2012).

Following golden era of antibiotic discovery, the efficiency of classical screening programs dropped significantly (Figure 1), making them unprofitable for pharmaceutical companies (Baltz, 2008). Due to the high rediscovery rate, it was hypothesized that the biosynthetic potential of the traditional producers had been fully explored (Cooper & Shlaes, 2011, Hutchings *et al.*, 2019). Recent advancements in sequencing technologies have resulted in the availability of a repository of genome sequence information (Katz & Baltz, 2016, Baltz, 2008). Once thought to be largely devoid of new drugs, these technologies revealed evidence for a treasure trove of yet undiscovered molecules that lie hidden in the biosynthetically talented bacterial genera (Bentley *et al.*, 2002, Ikeda *et al.*, 2003, Wipat & Harwood, 1999). Next-generation sequencing of microbial genomes has rejuvenated the discovery and characterization of diverse natural products and revealed the still largely underexplored biosynthetic potential of bacteria (Bentley *et al.*, 2002).

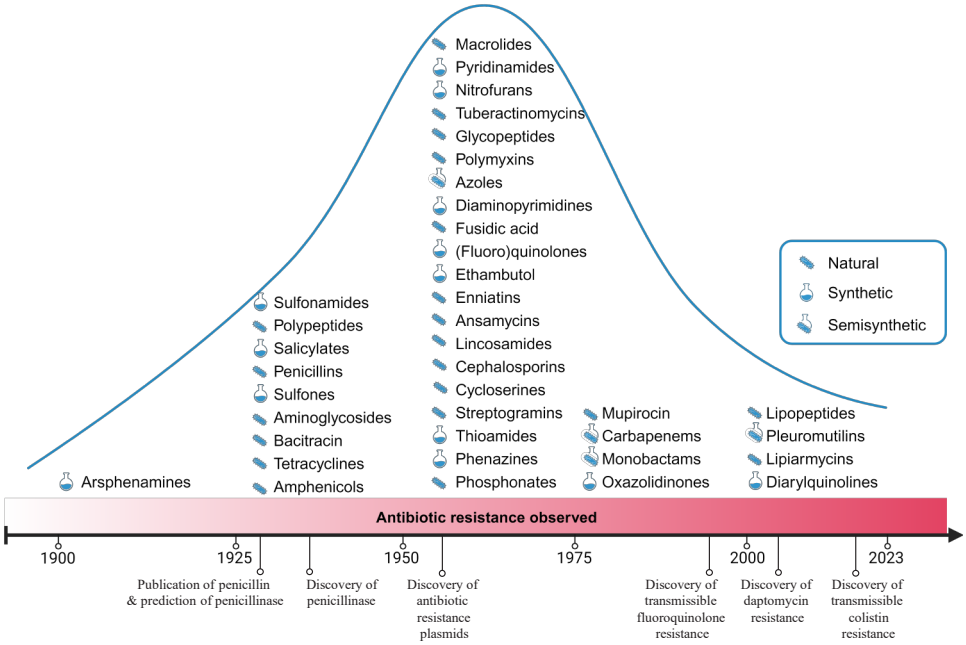

Figure 1. Timeline of antibiotic discovery and antibiotic resistance appearance. The icon marks the origin of the antibiotic class. The period between the 1950s and 1970s was the golden era of the discovery of novel antibiotic classes, but the development declined between the 1970s and the 1990s. Note that the increase in antibiotic resistance parallels a lack of success in discovering new antibacterial drugs.

Genome mining to evaluate the biosynthetic potential of bacteria

Thus, as the bacterial DNA ultimately encodes the production of all specialized metabolites, genome mining may offer a new route to natural product-based medicine discovery. In the bacterial genomes, the genes encoding biosynthetic pathways for natural products are usually organized in so-called biosynthetic gene clusters (BGCs) (Martin & Liras, 1989). It is efficient for microbes to have all the genes encoding a biosynthetic pathway in the same place; in this way, the entire production line can be activated by transcribing DNA from this genomic region. The toolkit for identifying BGCs involves methods to pinpoint key microbial sources, obtain high-quality genome data through sequencing technologies like Illumina, PacBio, and Oxford Nanopore, and analyze it using diverse bioinformatic tools (Mullowney *et al.*, 2023). Genome mining tools antiSMASH (Blin *et al.*, 2023) and PRISM (Skinnider *et al.*, 2020)2020 have been developed to identify and annotate microbial BGCs. With the help of antiSMASH (Blin *et al.*, 2023) we can identify BGCs in microbial genomes and classify them based on the type of molecule they are predicted to produce, such as polyketides (Ray & Moore, 2016, Donadio *et al.*, 2007, Shen, 2003), ribosomally synthesized and post-translationally modified peptides (RiPP) (Montalban-Lopez *et al.*, 2021), terpenes (Avalos *et al.*, 2022, Schulz &

Dickschat, 2007), and nonribosomal peptides (Sieber & Marahiel, 2005, Nivina *et al.*, 2019, Süssmuth & Mainz, 2017). PRISM uses pHMMs to detect BGCs and additionally predicts biological activity through machine learning (Skinnider *et al.*, 2020)2020. Alternative deep-learning genome mining software for BGC prediction, such as ClusterFinder (Cimermancic *et al.*, 2014), DeepBGC (Hannigan *et al.*, 2019), GECCO (Carroll *et al.*, 2021), and SanntiS (Sanchez *et al.*, 2023), uncovered undetectable putative BGCs that may code for natural products with novel bioactivities. BAGEL (van Heel *et al.*, 2013) and RODEO (Tietz *et al.*, 2017) are two genome mining tools designed explicitly for genome mining of RiPPs. Additionally, RiPPs may also be identified by their RiPP recognition Element (RRE), and the recently developed RRE-finder was incorporated into the shell of antiSMASH 6 (Kloosterman *et al.*, 2020b). Nevertheless, the diversity of RiPPs is vast, and the chemical space is expanding rapidly. To allow their detection, genome mining algorithms were developed to identify BGCs not captured using canonical rule-based annotation approaches (de Los Santos, 2019, Kloosterman *et al.*, 2020a, Merwin *et al.*, 2020, Tietz *et al.*, 2017). This includes the machine learning-based decRiPPter algorithm, which allowed the identification of a novel subclass of lanthipeptides that had escaped identification via existing algorithms (Kloosterman *et al.*, 2020a). The extensive list of the genome-mining software, tools, and databases utilized for NP discovery is available on the Secondary Metabolite Bioinformatics Portal (SMBP) (Weber & Kim, 2016).

The immense amount of genome data has prompted genome mining analyses conducted on a large scale. Rather than mining a single genome at a time, we can now incorporate big data approaches to mine a large pool of genomes or metagenomes for novel NPs (Medema & Fischbach, 2015). This approach is called "global genome mining" (Medema *et al.*, 2021). Bioinformatics tools that facilitate the analysis of big BGC datasets include the EvoMining (Sélem-Mojica *et al.*, 2019), ARTS (Alanjary *et al.*, 2017), BiG-SCAPE and the CORASON (Navarro-Muñoz *et al.*, 2020). Moreover, a highly scalable tool BiG-SLiCE was recently developed to deal with extensive data from global genome mining (Kautsar *et al.*, 2021b).

With the simultaneous sequencing of hundreds to thousands of microbial genomes becoming more common, the large quantities of BGC data resulting from this pose an opportunity and challenge. Databases such as antiSMASH-DB (Blin *et al.*, 2024)2024, IMG-ABC (Palaniappan *et al.*, 2020), and MIBiG (Terlouw *et al.*, 2023) play a critical role in analyzing BGCs by enabling the comparison of newly sequenced BGCs with those predicted and experimentally characterized in the past. In addition to the mentioned databased for exploring BGCs from the publicly available data, the recently established BiG-FAM database focuses on the gene cluster family (GCF) relationship of the detected BGCs and enables GCF-

based exploration and homology searching of >1.2 million BGCs harbored by >200,000 microbial genomes (Kautsar *et al.*, 2021a).

Translating the genetic code into compounds

Once BGCs of interest have been identified based on compound class, resistance genes, or phylogeny, predicting the specific final product remains challenging (Scherlach & Hertweck, 2021). For some natural product classes, we have a relatively good understanding of how the encoded biosynthetic machinery builds a natural product, and thus, we can attempt to predict the scaffold structure of the produced natural product. Among all natural product classes, the chemical structure can be predicted for NRPSs, modular polyketide synthase (PKSs), and RiPPs (Scherlach & Hertweck, 2021). The core structure and post-translational modifications (PTMs) of the RiPPs can be reliably predicted by RODEO (Tietz *et al.*, 2017), RiPPER (Moffat *et al.*, 2021), antiSMASH (Blin *et al.*, 2023) and PRISM (Skinnider *et al.*, 2020)2020.

Considering the scope of the thesis, the prediction principle will be illustrated based on the NRPS assembly line. NRPS machineries are complex, multi-domain enzymes that select and condensate amino acids iteratively to build up peptidic natural products (Sieber & Marahiel, 2005, Nivina *et al.*, 2019, Süssmuth & Mainz, 2017). To achieve nonribosomal peptide synthesis, minimal enzymatic complexes require four distinct domains called core or essential domains: condensation (C), adenylation (A), peptidyl carrier protein (PCP), and a terminal thioesterase (TE) domain (Figure 2) (Marahiel, 2016). Three domains make up the minimum elongation module, with C domains catalyzing condensation, A domains mediating building block selection through adenylation and subsequent thiolation, and non-catalytic PCP domains functioning as mechanical arms to which intermediate scaffolds are tethered between reactions (Winn *et al.*, 2016). An NRPS assembly line only requires a single TE domain, typically at the end of an NRPS enzyme, which is responsible for chain release (Mootz *et al.*, 2002).

As A domains govern building block selection, many NRP structure prediction algorithms attempt to predict the selected substrate from A domain sequence. Examples of such machine learning algorithms include NRPSpredictor2 (Röttig *et al.*, 2011), AdenylPred (Mongia *et al.*, 2023), SANDPUMA (Chevrette *et al.*, 2017), and AdenPredictor (Chevrette *et al.*, 2017), which were trained on A domain sequences of known specificity.

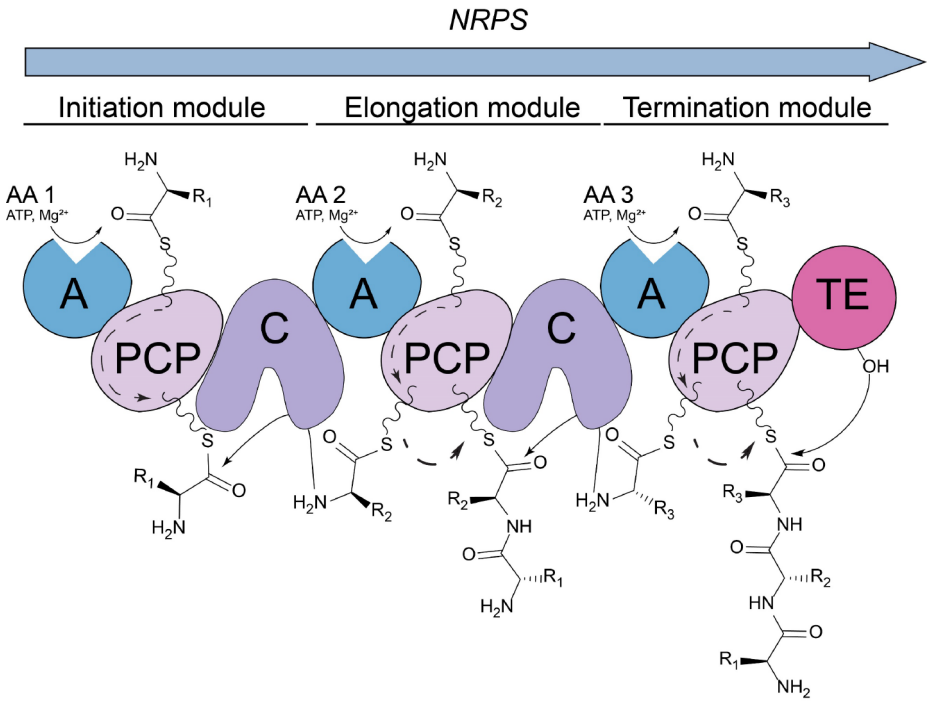

Figure 2. Schematic representation of NRP biosynthesis. The NRP biosynthesis involves the activation of amino acid substrates via ATP, facilitated by the adenylation domain (A). Afterward, the aminoacyl-AMP intermediate is captured by the thiol group of the flexible 4'-phosphopantetheine arm linked to a peptidyl carrier protein (PCP) domain. Condensation domains (C) subsequently facilitate peptide bond formation between thioester intermediates loaded onto neighboring PCP domains. Lastly, the final module hosts an additional thioesterase domain (TE), which is responsible for either hydrolysis or cyclization to release the product from the NRPS (adapted from Winn *et al.*, 2016).

However, the diversity and complexity of NRPs have made systematic investigation difficult, and consequently, most genetically encoded NRPs in bacteria have been overlooked. An alternative way to investigate the potential of novel scaffolds predicted by genome mining is via organic synthesis or chemoenzymatic total synthesis. The approach of synthetic bioinformatic natural product (syn-BNP) involves bioinformatic prediction of the natural product structure based on the genome sequence with subsequent organic synthesis of compounds, circumventing the need for bacterial culture and gene expression (Scherlach & Hertweck, 2021, Chu *et al.*, 2020). Concurrently, solid phase peptide synthesis (SPPS) of structurally diverse peptides has become rapid and economical, making NRPS gene clusters ideal candidates for a syn-BNP approach. This new methodology completely avoids the issues of isolation and scalability of target compounds. The syn-BNP approach was successfully applied to discover new antimicrobial entities from sequenced bacterial genomes or metagenomes, with an emphasis on the de novo discovery of novel clinically relevant NRP antibiotics. This approach led to the discovery of humimycin A and B (Chu *et al.*, 2016)2016,

kanglemycins (Peek *et al.*, 2018), macolacin (Wang *et al.*, 2022c), cilagicin (Wang *et al.*, 2022b) with potent bioactivity against multidrug-resistant (MDR) pathogens.

Examples of clinically used nonribosomal peptides

NRPs are among the most widespread groups of natural products with diverse chemical structures (Wang *et al.*, 2014). NRPs often contain both proteinogenic and nonproteinogenic amino acids, significantly expanding their chemical diversity and bioactivity spectrum (Caboche *et al.*, 2010). The rising issue of antibiotic resistance has led to increased interest in the therapeutic potential of NRP antibiotics (Hancock & Chapple, 1999). Peptide antibiotics currently used in clinics, such as bacitracin, daptomycin, polymyxin, and vancomycin constitute effective treatments for infections caused by multidrug-resistant pathogens (Figure 3) (Felnagle *et al.*, 2008, Liu *et al.*, 2019).

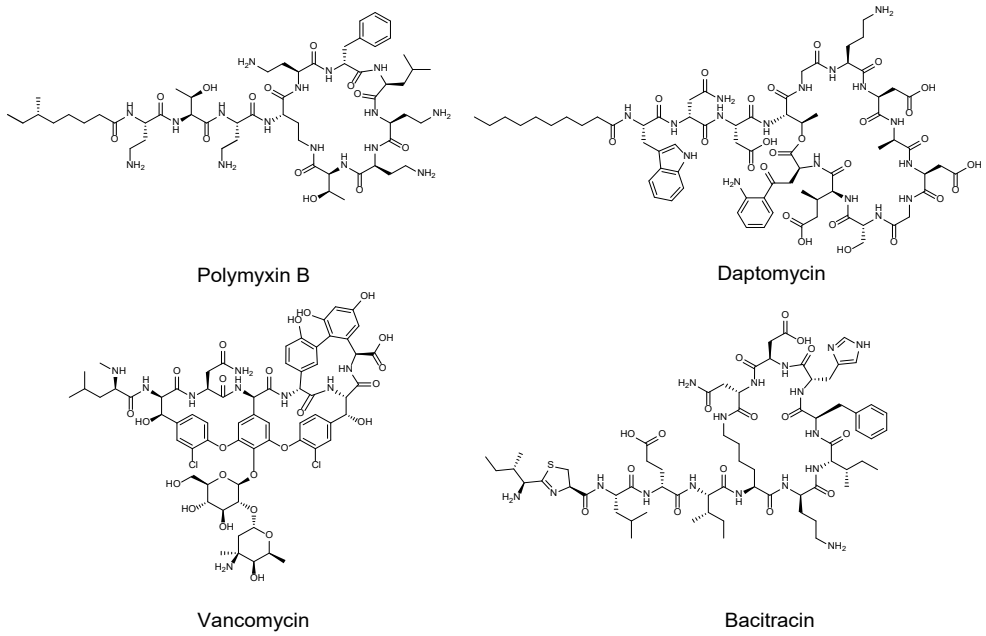

Figure 3. Structures of clinically used NRPs.

Polymyxins are active against Gram-negative bacteria and are typically only used to treat infections caused by multidrug-resistant organisms, including carbapenem-resistant *Enterobacterales*, *Acinetobacter baumannii,* and *Pseudomonas aeruginosa (Nang et al., 2021, Slingerland & Martin, 2024)*. The mode of action of polymyxins involves binding to the lipid A component of lipopolysaccharides (LPSs) in the outer membrane of Gram-negative bacteria. The positively charged 2,4-diaminobuturic acid residues of polymyxins electrostatically

interact with the negatively charged phosphate groups of lipid A, facilitating binding and leading to membrane disruption and increased permeability (Mohapatra *et al.*, 2021). In contrast, NRPs like bacitracin, vancomycin, and daptomycin are active preliminary against Gram-positive bacteria and act as inhibitors of cell wall biosynthesis at the different stages of this complex process (Buijs *et al.*, 2023). Daptomycin and vancomycin are considered "last resort" antibiotics for the treatment of severe Gram-positive infections (Gonzalez-Ruiz *et al.*, 2016, Griffith, 1981). Vancomycin targets and tightly binds lipid II, an essential building block of the cell wall, resulting in the cessation of cell wall biosynthesis (Williams & Butcher, 1981, Williams, 1984). The mode of action of daptomycin is multifaceted and involves the targeting and depolarization of bacterial cell membrane alongside the inhibition of cell wall biosynthesis by direct binding to lipid II (Kotsogianni *et al.*, 2021, Grein *et al.*, 2020). Other NRPs, like bacitracin, sequester the membrane-associated phospholipid undecaprenyl pyrophosphates (C_{55}PP), transport molecule that carries the building blocks of the peptidoglycan bacterial cell wall (Ming & Epperson, 2002). The binding prevents peptidoglycan synthesis, thereby inhibiting bacterial cell growth. The recent discovery of cilagicins, which exhibit a unique mode of action and bind both C_{55}PP and undecaprenyl phosphate (C_{55}P) involved in cell wall biosynthesis, highlights the significant untapped potential of nonribosomal peptides (NRPs) in combating antibiotic-resistant bacteria (Wang *et al.*, 2022b).

Exceptions to the rules of NRP biosynthesis

Still, translating genomic knowledge into molecules remains challenging. The diversity of NRPS-derived peptides is not only determined by the incorporation of different amino acid building blocks but also by an ever-increasing number of characterized tailoring reactions and enzymatic domains (McErlean *et al.*, 2019). Moreover, the promiscuity of NRPS modules allows the production of multiple compounds encoded by a single BGC (Duban *et al.*, 2022). In addition, a constantly expanding library of scientific literature unravels new examples of nonribosomal peptide synthetases exhibiting rare domains and noncanonical module organization, leading to unique microbial strategies of natural products biosynthesis (Chevrette *et al.*, 2020). Moreover, recent publications demonstrate a much greater mechanistic diversity in NRPS assembly lines than previously anticipated. Recent characterization of several NRPS megasynthases revealed examples of the loss of the co-linearity rule (Wenzel & Müller, 2005). It turned out that modules from bacterial NRPSs can be used repeatedly, skipped, or even split in the biosynthesis of a single product (Gewolb, 2002, Marahiel, 2009, Wenzel & Müller, 2005).

One notable example entails collaborative natural product biosynthesis, involving crosstalk between two separate BGCs. Such pairing of biosynthetic enzymes from distinct BGCs within the same biosynthetic pathway gives rise to increased chemical diversity

and an expanded scope of biological activity. Over the past few years, this phenomenon of collaborative biosynthesis has received increased attention, and specialized metabolites produced through such strategies are known as 'hybrid molecules' or 'chimeras' (Mevers *et al.*, 2019, Yin *et al.*, 2022). Chimeric biosynthesis has previously been observed in various NRPS and PKS (hybrid) systems, including the biosynthesis of azaphilone (Huang *et al.*, 2020), penilactones (Fan *et al.*, 2019), spirotryprostatin A (Yan *et al.*, 2019), and echinocandin B (Cacho *et al.*, 2012).

Despite the impact of genomics, it only delivers possible leads, which needs to be supported by experimental evidence provided by microbiology, analytical chemistry, and bioassays to purify and identify the NPs and determine their mode of action. Analytical chemistry techniques, such as high-resolution mass spectrometry and NMR (nuclear magnetic resonance spectroscopy), complement theoretical in silico predictions. They play a crucial role in bridging the gap between genomic potential and the actual production of metabolites (Caesar *et al.*, 2021).

Expanding chemical space and identification of bioactive natural products

Activation and characterization of cryptic BGCs

Many BGCs are not expressed during standard laboratory conditions, and thus, their cognate natural products escape attention (Nett *et al.*, 2009, Kolter & van Wezel, 2016). To unlock the full potential of bacteria, it's crucial to understand their ecological context, as this will provide clues to the activation mechanisms of their biosynthetic pathways (van Bergeijk *et al.*, 2020, Zhu *et al.*, 2014a). The signals that mediate the production of specialized metabolites are diverse and include physical cell-cell interactions, nutrient depletion, enzymatic conversion of precursors to active metabolites and microbial small molecules (Hoskisson & Fernández-Martínez, 2018, van der Heul *et al.*, 2018, Huang *et al.*, 2005). However, for many interactions, the signals and molecular mechanisms remain unknown. Several methods have been developed to boost the production of hidden natural products, including ribosome engineering, co-culture, varying culturing conditions, insertion of active promoters, and high-throughput elicitor screening (Bertrand *et al.*, 2014, Takano, 2006, Zhu *et al.*, 2014a, Bode *et al.*, 2002). The expression of bacterial BGCs can be manipulated by altering the culturing conditions. This is the basis of the one-strain–many-compounds (OSMAC) concept, meaning that a strain can produce a plethora of distinct compounds if exposed to different abiotic or biotic conditions (Bode *et al.*, 2002). In co-culturing, microbial partners are added to a single culture to trigger the expression of metabolic programs involved in defense or nutrient competition (Sugiyama *et al.*, 2015, Hoshino *et al.*, 2015, Bertrand *et al.*, 2014, Wu *et al.*, 2015b, Schroeckh *et al.*, 2009). As an alternative, high-throughput elicitor screening can be used to induce the expression of silent BGCs using the small compounds. This is an efficient platform that enables the discovery

of small molecule activators that lead to the induction of silent biosynthetic clusters as well as structural and functional elucidation of their products (Seyedsayamdost, 2014, Moon *et al.*, 2019a, Moon *et al.*, 2019b). Moreover, advances in synthetic biology enable the engineering of biological systems to express customized biosynthetic pathways. Synthetic biology tools can alter the genome of selected microorganisms to induce the expression of BGCs and the efficient production of natural products (Smanski *et al.*, 2016, Zhang *et al.*, 2019). The resulting "cell factory" provides improved genomic stability and allows the optimization of the targeted biosynthetic pathway in culturable and biochemically characterized systems.

LC-MS-based metabolomics methods for data acquisition

For metabolite profiling, culture extracts are analyzed by NMR spectroscopy, high-resolution mass spectrometry (HRMS), or combined methods involving upstream liquid chromatography (LC), such as LC-HRMS and more recent LC-NMR (Atanasov *et al.*, 2021, Wu *et al.*, 2015a). Compatibility of reversed-phase (RP) LC with biological samples allowed its utilization in combination with accurate mass MS (QToF or Orbitrap technologies) (Roux *et al.*, 2011). LC-MS established itself as the technology of choice for microbial metabolic profiling due to the widespread availability of suitable instrumentation and ease of both access and use, combined with good sensitivity, specificity, and resolving power. Considering the scope of the thesis, I will here focus mainly on LC-MS-based metabolomics. For NMR-based metabolomics the reader is referred to reviews elsewhere (Markley *et al.*, 2017, Wu *et al.*, 2016a, Grienke *et al.*, 2019), and likewise for reviews of MS-based approaches (Krug & Müller, 2014, Bouslimani *et al.*, 2014, Spraker *et al.*, 2020).

The key challenges scientists face when targeting the metabolome are analytical reproducibility, feature coverage, dynamic range, component annotation, data analysis, and processing large sample sets. LC-MS has enabled the detection of thousands of mass features within a single biological sample. Therefore, one of the main challenges in LC-MS-based metabolomics is the comprehensive and accurate analysis of large datasets. Freely available data preprocessing tools, such as XCMS (Smith *et al.*, 2006), Mzmine (Pluskal *et al.*, 2010), MetaboAnalyst (Pang *et al.*, 2022), MetAlign (Lommen, 2009), Metabolomic Analysis and Visualization Engine (MAVEN) (Clasquin *et al.*, 2012), MET-COFEA (Zhang *et al.*, 2014) as well as commercial software, have been developed to facilitate data processing. Moreover, the recently developed publicly available R-based web application Metabolomics Explorer (MetEx) enables users to quickly and intuitively analyze comprehensive metabolomics datasets. MetEx facilitates analysis of complex datasets, consisting of retention time, *m/z*, and MS intensity features, as a function of hundreds of conditions (Covington & Seyedsayamdost, 2021).

LC-MS-based metabolomics and the discovery of bioactive compounds

Metabolomics measures multiple metabolic responses in living systems, allowing simultaneous tracking of changes in hundreds of small molecules. However, the complexity and diversity of chemical structures make metabolomic analysis challenging. The methodologies employed to interpret high-throughput metabolomics data mainly derived from earlier emerging omics technologies. Classic approaches aim to assess group-wise differences, either in a univariate, parameter-by-parameter fashion, e.g., *t*-test, analysis of variance (ANOVA), or using multivariate techniques (e.g., MANOVA, ASCA, PCA, PLS) (Bartel *et al.*, 2013). Statistical models such as Pearson correlation, partial least squares (PLS), discriminant analysis (PCA-DA, PLS-DA, OPLS-DA), and hierarchical cluster analysis (HCA) are often used to link metabolite fingerprints to bioactivity data (Bartel *et al.*, 2013). The general idea behind supervised methods is to unravel distinct metabolite profiles that are associated with the observed bioactivity of the sample. A recent extension to the PLS repository is the orthogonal-PLS (OPLS) method. In the OPLS analysis, the data variation is split into the variance of interest and a noise part, which is unrelated to the response (Dettmer *et al.*, 2007). This leads to a simplified interpretability of the resulting components, allowing additional assessment of within- and between-group variance.

While the approaches mentioned above provide a list of the highlighted features, the information obtained is at the prediction level. Multiple analytical modules involving different bioassays and detection technologies should be linked to allow simultaneous bioactivity evaluation and identification of compounds in complex compound mixtures. Such approaches may help provide a more comprehensive insight into compounds responsible for the bioactivity of interest.

GNPS molecular networking and dereplication of natural products

In the pursuit of antibiotic discovery, a pivotal role is played by Global Natural Products Social molecular networking (GNPS) (Wang *et al.*, 2016). This open-access platform is a hub for sharing, processing, and visualizing tandem mass spectrometry data (Quinn *et al.*, 2017). The concept of molecular networking is based on the organization and visualization of tandem MS data through a spectral similarity map, revealing the presence of homologous MS^2 fragmentations. As structurally related compounds share similar fragmentation spectra, their nodes tend to connect and create spectral families (Wang *et al.*, 2016). Classical molecular networking has recently been integrated with feature detection methods to introduce Feature-Based Molecular Networking (FBMN). This approach enables the discrimination of isomers within the molecular network and incorporates quantitative information generated by the feature detection tools (Nothias *et al.*, 2020).

Molecular networking lays the groundwork for bioactivity-based molecular networking. Combining the MN with metabolite "bioactivity scores" enables the detection and relative quantification of LC-MS/MS spectral features across chromatographic fractions. The "bioactivity scores" are calculated based on the Pearson correlation between feature intensity across samples and the bioactivity level associated with each sample (Nothias *et al.*, 2018). The correlation of the metabolite quantity and observed bioactivity facilitates the discovery of biologically active compounds within complex metabolomic datasets.

The major challenge in metabolomics is the identification of previously reported compounds, known as structural dereplication (Gaudencio & Pereira, 2015, Hubert *et al.*, 2017, Silver, 2011). Dereplication approaches include the determination of molecular formula using the Seven Golden Rules, Sirius 2, and MS-FINDER software and searching for compounds described by this formula in chemical structure databases (Gaudencio & Pereira, 2015, Mohamed *et al.*, 2016). The primary challenge of these approachs is the exponential increase in the number of potential formulas as the molecular mass of metabolites grows. Furthermore, existing chemical databases frequently contain numerous compounds with identical formulas, which complicates the annotation process.

Currently, the combination of molecular networking with annotation methods such as MS2LDA (van der Hooft *et al.*, 2016), Network Annotation Propagation (NAP) (da Silva *et al.*, 2018) or DEREPLICATOR (Mohimani *et al.*, 2018) accelerates the identification of unannotated ions. MolNetEnhancer (Ernst *et al.*, 2019) combines outputs from MS2LDA, NAP, DEREPLICATOR, and molecular networking, along with the automated chemical classification from ClassyFire (Djoumbou Feunang *et al.*, 2016), to assign structural features to chemical classes. Algorithms such as DEREPLICATOR or recently published VarQuest (Gurevich *et al.*, 2018) were designed to expand the library search to identify variants of known peptidic natural products. The newly developed tool Qemistree is employed to classify MS data into a tree structure (Tripathi *et al.*, 2021).

Future perspectives

A number of innovative methods have evolved to enhance the synergy between genomics and metabolomics for natural product discovery. Future developments for the processing of metabolome data will be focused more on continuous platform improvement, elucidation of structural identification, and functional interpretation of metabolomic data. As multi-omics technologies mature, there is an increasing need for data integration between platforms. One of the key challenges in the field of omics-based natural product discovery is the set-up of high-quality datasets to train deep learning algorithms and appropriate strategies for algorithm validation.

Despite the advances in the genomics and metabolomics fields, the current state of the art regarding BGC and natural product diversity limits our ability to guide drug discovery.

It is, therefore, necessary to further characterize the extant microbial diversity, pathways, and regulatory mechanisms for natural product biosynthesis. Moreover, characterizing the functions and modes of action of natural products in native microbial communities will be crucial for prioritizing BGCs and eliciting their expression. This should bring many more natural products to light, potentially contributing to future drug candidates.