Discovery of gene sub-clusters encoding natural product substructures

MSc Thesis

Joris Louwen, 960516530090

04/10/19

Chair group of Bioinformatics, Wageningen University

Supervisors:

Justin van der Hooft, Satria Kautsar, Marnix Medema

Content

Abstract 2
Introduction 3
Methods 4
Data selection and processing 5
Sub-cluster detection: re-implementation of the statistical method 5
Sub-cluster detection: LDA 6
Results 6
Detecting sub-clusters with the previously established statistical method 6
LDA as a new tool for sub-cluster detection 7
Both methods capture the majority of experimentally validated sub-clusters 8
Exploring the sub-cluster motifs 8
Identifying the BGC for heronapyrrole biosynthesis 9
Correlation analysis 10
Discussion 11
References 13
Supplementary data 14
Supplementary methods 14
Figures 17
Tables 20
Supplementary files 23

Abstract

Bacterial specialised metabolites are a rich source of natural products (NPs). The genes responsible for their biosynthesis are physically clustered on the genome in biosynthetic gene clusters (BGCs). Many BGCs consist of multiple groups of co-evolving genes called sub-clusters, each of which is responsible for synthesising a specific chemical moiety in the NP. Sub-clusters therefore provide an important link between the substructures of an NP and its BGC, highlighting the importance of sub-clusters for structural predictions. Here, we improved the existing method for sub-cluster detection by making it more scalable, reducing redundancy and using the antiSMASH database as a data source, which constitutes a ten-fold increase in training data. Additionally, we introduced a text-mining algorithm called Latent Dirichlet Allocation (LDA) as a novel unsupervised method for sub-cluster detection. LDA identifies groups of co-occurring genes in the data and clusters them into sub-cluster motifs. Using LDA, we were able to identify 71% of the experimentally validated sub-clusters from the SubClusterBlast module in antiSMASH. Furthermore, we annotated 50 sub-cluster motifs with structural information. This could readily be used by BGC prediction tools like antiSMASH to enhance structure prediction and to provide novel structural information to unclassified BGCs. We showed a direct application of our method by proposing the BGC for heronapyrrole biosynthesis. Finally, we used a systematic approach to link sub-clusters to substructures, which could be used in the future to connect BGCs to their NPs in an automated manner.

Introduction

A considerable part of bacterial metabolism is dedicated to the synthesis of specialised metabolites. While these molecules govern intra-and interspecies interactions in general, they are bioactive compounds used for defence, making them invaluable for bacterial survival (Traxler *et al.*, 2015). Specialised metabolites are natural products (NPs) with many uses in pharmaceutical, agricultural and dietary agents, like antibiotics, antitumor agents and herbicides (Dayan *et al.*, 2009; Li *et al.*, 2009). NPs consist of a spectrum of different chemical classes, which are often highly complex in structure. Intriguingly, the genes necessary for the synthesis of NPs cluster together physically in biosynthetic gene clusters (BGCs) (Medema *et al.*, 2014). The search and discovery of new BGCs is thus crucial for identifying new NPs, which is especially important in the field of antibiotics as antibiotic resistant bacteria become more prevalent (Chevrette *et al.*, 2018).

Traditionally, novel NPs were identified using low-throughput wet lab experiments (Katz *et al.*, 2016). In recent years however, genome mining approaches have become increasingly interesting for NP discovery, due to the growing availability of genomic data. Multiple algorithms exist that search bacterial genomes for putative BGCs, such as antiSMASH, ClusterFinder and PRISM (Kai Blin *et al.*, 2019b; Cimermancic *et al.*, 2014; Skinnider *et al.*, 2017). These methods have provided a better understanding of BGC diversity and the evolutionary mechanisms that govern BGC diversity. Even using conservative constraints, Cimermancic *et al.* (2014) estimated that over 6,000 broad BGC families exist, each of which consists of multiple BGCs that synthesise distinct molecules from a common scaffold. Rapid evolution seems to be an important factor for the large BGC diversity. This shows from the high frequency of horizontal gene transfer and the high rates of insertions, deletions, duplications, and rearrangements that BGCs exhibit (Medema *et al.*, 2014). Another striking observation is the modular structure that several types of BGCs display. In particular, BGCs encoding for polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) seem to consist of multiple modules or sub-clusters of co-evolving genes, each responsible for synthesising specific chemical moieties (Del Carratore *et al.*, 2019; Fischbach *et al.*, 2008; Medema *et al.*, 2014). Sub-clusters therefore provide a direct link between the substructures of an NP and its BGC. This makes information about sub-clusters and the substructures they synthesise highly valuable for structure prediction, which would be a great asset for tools like antiSMASH. Not only would it improve the structure prediction of existing BGC classes but it would also be possible to predict structures of currently unclassified BGCs, such as the 10,000 unclassified BGCs in the antiSMASH database (K. Blin *et al.*, 2019a). In synthetic biology, the notion of BGC modularity has already been used to generate novel compounds by recombining core genes from PKSs and NRPSs, and by altering single modifying genes like methyltransferases (Kim *et al.*, 2015; Menzella *et al.*, 2005). However, insight into sub-clusters allows the possibility to interchange complete chemical moieties instead of the core structure or small modifications. In addition to synthetic biology, BGC modularity poses a great opportunity to connect metabolomics experiments to sub-cluster data. Chemical moieties identified from fragments in mass spectrometry (MS) data could be linked to sub-clusters responsible for their synthesis, which could lead to MS-guided genome mining (Del Carratore *et al.*, 2019). Recent advances in substructure modelling aid such metabologenomic approaches (van der Hooft *et al.*, 2016). More knowledge about sub-clusters is thus beneficial for understanding BGC evolution, and has many practical applications.

Recently, Del Carratore *et al.* (2019) developed a method for the detection of sub-clusters in BGCs. They constructed Clusters of Orthologous Groups (COGs) and used a statistical approach to determine if a group of co-occurring COGs represents a sub-cluster. Subsequently, they made a score to be able to rank the sub-clusters based on relevance and reliability. With this method, 185,718 putative sub-clusters were found in a set of 12,842 BGCs, which were predicted with antiSMASH. Well characterised sub-clusters were found to be present in the set of sub-clusters with high scores, demonstrating the effectiveness of their method. However, the number of sub-clusters found in this analysis is very large. This can partly be explained by the nested structure of a lot of sub-clusters, in which smaller, less specific sub-clusters are contained in larger, more specific sub-clusters. Another reason for the inflated number of detected sub-cluster might be that phylogenetic bias is not considered. As such, the presence of many highly similar

BGCs could result in the detection of artificial sub-clusters that consist of almost entire BGCs. Additionally, this method is not very scalable as it relies on extensive all-vs-all searches to construct the COGs.

Here, we constructed an improved method for sub-cluster detection that is scalable to large datasets and takes phylogenetic bias into account by filtering the input in a more advanced way. Scalability is achieved by tokenising BGCs into strings of Pfam domains as a proxy for sequence identity, instead of extensive COG construction. As Pfams are quite broad sequence models, we increased the resolution by splitting the most important Pfams into a number of subPfams, each of which is a more narrow domain model that covers a subset of a Pfam's sequence space. To improve the previous statistical method, we removed nested sub-clusters, collapsed similar sub-clusters into families, and similar families into clans. In addition to the previous method, we implemented a text mining algorithm, Latent Dirichlet Allocation (LDA), for the purpose of sub-cluster detection (Blei *et al.*, 2003). LDA has already been used successfully in other branches of natural science, such as genomics and metabolomics (Chen *et al.*, 2010; van der Hooft *et al.*, 2016). In our case, LDA learns a set of sub-cluster motifs from a collection of BGCs, which are in turn used to infer multiple sub-clusters in a BGC. We applied our pipeline to the antiSMASH database, which is a considerable improvement in comparison with the previous method as it contains over ten times as many BGCs (150,000) from almost 25,000 bacterial species that are chosen to reduce taxonomical bias (K. Blin *et al.*, 2019a). For the validation of our method we included the MiBIG database, a set of curated BGCs with structural information, in which a number of sub-clusters have been experimentally verified (Medema *et al.*, 2015). With our approach, we were able to characterise 50 different sub-cluster motifs present in diverse BGC classes. The remaining 950 sub-cluster motifs remain largely unexplored, of which many are likely to encode useful substructures. Using one of the characterised sub-cluster motifs, we showed a direct practical application of our method by proposing the putative BGC for heronapyrrole production. Moreover, we could correlate sub-clusters to substructures in a systematic manner by including the Crüsemann dataset in which substructure models were created previously using the MS2LDA tool (Crüsemann *et al.*, 2017; Ernst *et al.*, 2019).

Methods

Python was used for all analyses in this project. Code is available at https://git.wageningenur.nl/louwe015/scripts-thesis. In our pipeline, each BGC is tokenised by converting all genes into strings of (sub)Pfam combinations (Figure 1). With a graph based filtering step redundant BGCs are removed from the dataset, after which we detect sub-clusters using two algorithms: a statistical method and Latent Dirichlet Allocation (LDA). The resulting sub-clusters of both methods are annotated with substructures and can be used to predict sub-structures in BGCs. These steps are described in the following sections, while the supplementary methods provide more detailed explanations.

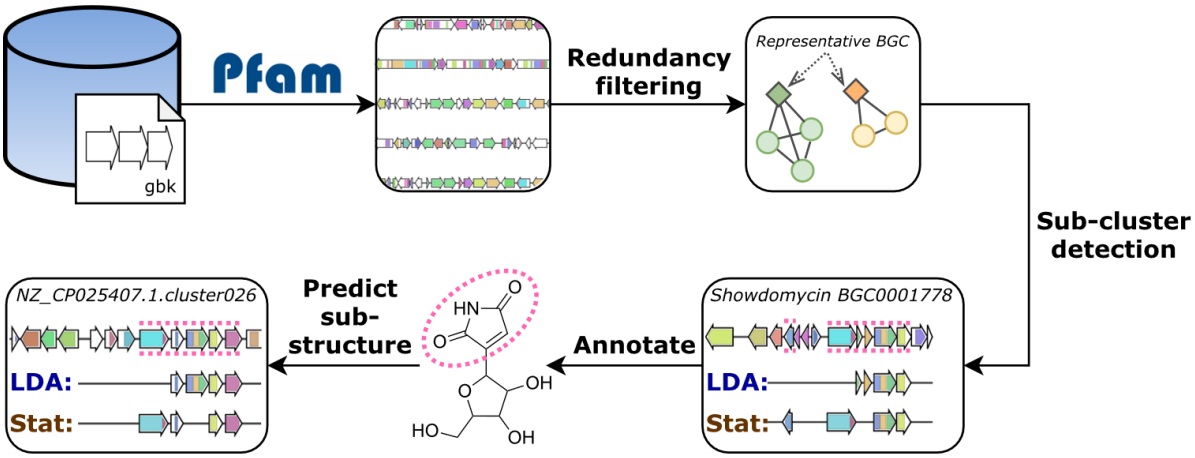

Figure 1 Workflow for the detection of sub-clusters. All genes in BGCs are converted into strings of Pfam domains, after which redundant BGCs are filtered out based on an Adjacency Index of domains. Sub-clusters are detected using two methods: Latent Dirichlet Allocation (LDA) and a statistical analysis (stat). BGCs from the MiBIG

database are used to annotate putative sub-clusters with sub-structures. These annotations are used to predict sub-structures in unknown BGCs.

Data selection and processing

The main dataset consisted of three data sources: the MiBIG database, the Crüsemann dataset and the antiSMASH database. Version 1.4 of the MiBIG database was used which contains 1,819 BGCs (https://dl.secondarymetabolites.org/mibig/mibig_gbk_1.4.tar.gz). The Crüsemann dataset consists of 5,927 BGCs that originate from the 146 *Streptomyces* and *Salinispora* strains investigated by Crüsemann *et al.* (2017). antiSMASH 3.0 was used for the detection of BGCs in the Crüsemann dataset. The antiSMASH database (https://antismash-db.secondarymetabolites.org/) is comprised of 152,122 BGCs detected with antiSMASH 4.0 (Table S1). Additionally, the MiBIG database and the Crüsemann dataset were used together as the small dataset. BGCs were discarded if they were flagged by antiSMASH as laying on a contig-edge, as these BGCs are probably incomplete and less accurate. Additionally, BGC-class information was included in the analysis, by using the predicted antiSMASH classes for the antiSMASH database, and the output of a BiG-SCAPE run for the Crüsemann and MiBIG datasets (van der Hooft, J. J., personal communication).

BGCs were tokenised by converting each gene into a string of (sub)Pfam domains. To detect (sub)Pfams, the HMMER3 tool hmmscan was used with a custom profile hidden Markov model (pHMM) database consisting of Pfam database version 32.0 where 112 Pfams were replaced by corresponding subPfams (Bateman *et al.*, 2018; Mistry *et al.*, 2013). These 112 Pfams were selected as they are the most abundant biosynthetic Pfams in the antiSMASH database (Kautsar, S. A., personal communication, Supplementary files). To create subPfams, the multiple sequence alignment of a Pfam is split into clades, after which a new pHMM is built for each clade, each of which constitutes a subPfam (Figure S6A).

Redundant BGCs were removed from the analysis using a similarity network of BGCs, where BGCs were connected based on an Adjacency Index of domains higher than 0.95 or if BGCs were fully contained within one another. From each maximal clique in the network, only the BGC with the most domains was chosen to remain in the analysis (Table S1;Figure S7). After redundancy filtering, all non-biosynthetic domains were removed from all BGCs. To select biosynthetic domains, EC-associated Pfams were collected with ECDomainMiner, from which Pfams were selected if they occurred in pre-calculated BGCs (Alborzi *et al.,* 2017; Kautsar, S. A., personal communication). After manual curation this resulted in a list of 1,839 biosynthetic Pfams (Supplementary files). Additionally, Pfams that occurred less than three times in a dataset were removed as well as BGCs that contained less than two non-empty genes.

Sub-cluster detection: re-implementation of the statistical method

The statistical method for sub-cluster detection was re-implemented in python according to Del Carratore *et al.* (2019). Instead of representing genes as COGs as in the previous method, we represent each gene as a combination of its domains. First, all possible adjacency and co-localisation interactions between each pair of genes are counted. To assess whether an observed interaction between two genes occurs more than by random chance, one needs to distribute such a pair of genes randomly through the dataset and calculate the probability of the observed interaction. To reduce the computational burden of a permutation-based approach, for each pair of genes one gene is kept fixed while the other is being randomly distributed throughout the data. For an adjacency interaction this gives a hypergeometric equation describing all available positions of one gene while the other is fixed (Table S2;Equation 1). This follows from the fact that there are three options for the position of gene B while keeping gene A fixed: not adjacent to gene A (B_1), adjacent to gene A (B_2), or adjacent to gene A on both sides (B_3). N_1, N_2 and N_3 represent all available positions in these three categories, while N_{tot} represent all positions and B_{tot} all occurrences of gene B. For a co-localisation interaction the same applies, except for the fact that gene B can be co-localised with n_{max} genes A, where n_{max} is the number of genes A co-localised with gene B (Table S2;Equation 2). When n_{max} is large this becomes computationally hard, which is why we replaced duplicate genes with an empty gene (a dash) and placed one copy of the duplicate gene at the end of the cluster separated by an empty gene. This simplifies the equation as only two types of co-localisation need to be counted: co-localisation and no co-localisation (Table S2;Equation 3). A p-value can be calculated by summing all probabilities in the

hypergeometric distribution that correspond to a number of interactions higher or equal to the observed number of interactions. Or, to make it easier, by subtracting the sum of all possible interactions smaller than the observed interaction from one (Table S2;Equation 4).

Sub-cluster detection: LDA

Latent Dirichlet Allocation (LDA) is an unsupervised algorithm based on Bayesian probabilities, which we use to infer latent sub-cluster composition in BGCs (Blei *et al.*, 2003). It assumes a bag-of-words representation, where each BGC is depicted as a frequency vector of its domain combinations, not taking gene order into account. Analogous to topic modelling of text, BGCs, domain-combinations and sub-cluster motifs resemble documents, words and topics, respectively. The domain-combinations in each BGC are used to approximate the latent parameters: the sub-cluster motif distributions in the BGCs and the domain-combination distributions in the sub-cluster motifs. To do so, the number of sub-cluster motifs N has to be predefined, as well as hyperparameters α and β. We chose all prior parameters for each dataset specifically, where we chose N with the highest overlap with SubClusterBlast (Supplementary methods) and symmetric α=β=1/N. For the actual inference of the latent parameters, online variational Bayes was used, which is implemented as a multicore version in Gensim (Hoffman *et al.*, 2010; Rehurek *et al.*, 2010). In this implementation an LDA model is trained by updating it with mini-batches from the dataset, which has low time and memory complexity. We chose the chunksize of each mini-batch to be 5% of the dataset with a minimum chunksize of 2,000 for small datasets, which is loosely based on testing different chunksizes by Hoffman *et al.* (2010). We considered that using 2,000 iterations to train a model was sufficient after assessing that the log-likelihood converged sufficiently (Figure S8).

Each sub-cluster motif in an LDA model consists of a probability vector of domain combinations, representing the contribution of each domain combination to a sub-cluster motif. To filter out noise, we sorted this vector from high to low probability, summed the probabilities and included all domain combinations until 0.95 was reached. When a group of genes from a BGC match to a sub-cluster motif, this putative sub-cluster is assigned a probability representing how much of the BGC matches to this sub-cluster motif. We set a cut-off on the match probability of 0.05, which loosely corresponds to a BGC of 40 genes matching with at least 2 genes to a sub-cluster motif. Each gene in a match is also assigned a probability describing how well it fits in a match, for which we set a cut-off of 0.3. Lastly, we calculated an overlap score for each match, which we calculated by summing the domain combination probabilities from the sub-cluster motif present in the match. We set a quite liberal threshold of 0.15 on the overlap score, as this was the highest threshold that did not remove manually validated SubClusterBlast sub-clusters from the analysis.

Results

As our main dataset, we used our three data sources grouped together, the antiSMASH database, the MiBIG database and the Crüsemann dataset. Additionally, we used the MiBIG database and the Crüsemann separately as a small dataset, to see how sub-cluster detection changes when increasing the amount of data. We processed each BGC for sub-cluster detection by tokenising each gene in a BGC as a combination of (sub)Pfam domains, and performing redundancy filtering (Figure S6;Table S1). After these processing steps, the main datasets contained 60,028 BGCs with 10,539 domain combinations, while the small dataset contained 2,923 BGCs with 1,874 domain combinations.

Detecting sub-clusters with the previously established statistical method

The statistical method finds groups of genes that are either adjacent to each other, or co-localise in more BGCs than you would expect by random chance, reasoning that such a group of genes is a sub-cluster of co-evolving genes. Using this method we found 243,246 sub-clusters in the main dataset, and 15,798 sub-clusters in the small dataset. For both datasets, over 70% of the statistical sub-clusters contain less than ten genes, and a good portion of the sub-clusters occur in more than 10 BGCs, *i.e.* 14% in the main dataset and 7% in the small dataset (Figure S9).

For the main dataset, we found an average of 4 sub-clusters per BGC, while the previous approach resulted in around 14 sub-clusters per BGC. It therefore seems like the nested nature of the sub-clusters has decreased by performing filtering for redundant BGCs. However, looking at the statistical sub-clusters nested structures were still apparent. We therefore aimed to cluster nested and related sub-clusters together, which also provides more comprehensibility to the statistical sub-clusters. We performed two rounds of K-means clustering, in which we first clustered the statistical sub-clusters into 10,000 sub-cluster families (SCFs) and then clustered these SCFs into 2,000 sub-cluster clans (SCCs) using the SCF cluster centres. As an additional measure for reducing nested sub-clusters, we removed redundant sub-clusters in each SCF if they had the same occurrence as some bigger sub-cluster containing the redundant sub-cluster completely. This removed over half of the sub-clusters resulting in 108,085 sub-clusters. Although some SCCs grouped seemingly unrelated sub-clusters together that share only one gene, the majority of 1626 SCCs provide groups of related sub-clusters, sharing at least three genes. With these two simple steps we managed to improve the comprehensibility of the statistical sub-clusters drastically.

LDA as a novel tool for sub-cluster detection

In order to enrich the discovery of sub-clusters we present a new unsupervised method for sub-cluster detection with Latent Dirichlet Allocation (LDA). LDA is a Bayesian probabilistic model used to model topics of co-occurring words in text documents (Blei *et al.*, 2003). In an LDA model, each document is depicted as a mixture over latent topics, in which a topic is a distribution over the words present in the documents. In our case a document is a BGC, a word is a gene represented as a domain combination, and a topic can be thought of as a sub-cluster motif. This highlights the use of LDA for sub-cluster detection as we assume that a BGC is a combination of multiple different sub-clusters, which consist of co-evolving genes that co-occur in multiple BGCs. Another benefit of LDA is illustrated by the fact that a topic or sub-cluster motif has the potential to contain a set of core genes that synthesise the base of a sub-structure, along with additional modifying genes, hereby capturing sub-structure diversity. We constructed two LDA models, one with 1,000 sub-cluster motifs for the main dataset and one with 100 sub-cluster motifs for the small dataset, after which the main and small dataset were queried on their respective LDA models. In the main dataset, we identified around 250,000 sub-clusters, where each sub-cluster is a group of genes matching against a sub-cluster motif. Over 80% of the BGCs in the main dataset contained at least one sub-cluster motif (Figure S10). Many of the sub-clusters were uninformative as they contained only one gene from a sub-cluster motif, or sub-clusters and their motifs encompassing entire BGCs (Figure 2A-B). For a sub-cluster to be interesting we would expect its size to be between 2-12 genes, as experimentally characterised sub-clusters fall in this range. Many sub-clusters were of this expected size making these sub-clusters and their motifs interesting. As such, two experimentally verified sub-clusters of macbecin for methoxymalonate and AHBA provide an example as we were able to identify them in sub-cluster motifs 563 and 742, respectively (Figure 2C).

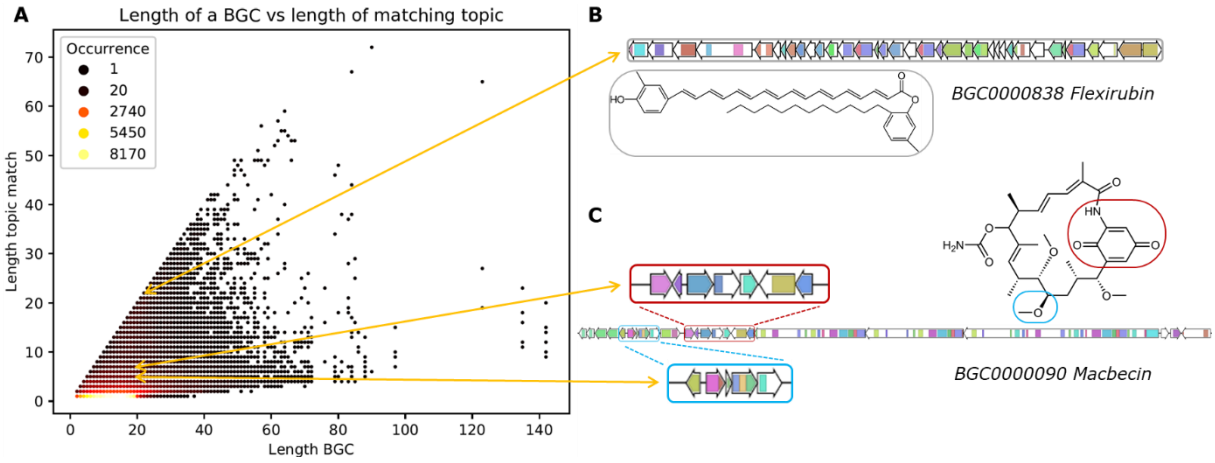

Figure 2 (A) Scatterplot of the length of each BGC (number of non-empty genes) from the main dataset versus the length of a match to a topic or sub-cluster motif, representing a sub-cluster. The colour of each dot indicates how many times a BGC with a certain length contains a sub-cluster with a certain length. (B) BGC for flexirubin where the identified sub-cluster found encompasses the entire BGC, demonstrating an uninformative result. (C) BGC for macbecin where the two characterised sub-clusters for AHBA (red) and methoxymalonyl (blue) are highlighted in the structure of macbecin (Zhang et al., 2008). Sub-clusters from (B) and (C) are linked to their corresponding location in (A).

Both methods capture the majority of experimentally validated sub-clusters

In order to validate the sub-cluster detection methods, we used a set of 109 experimentally validated sub-clusters. These 109 known sub-clusters are the only centrally stored validated sub-clusters, which are used by the SubClusterBlast tool in the antiSMASH framework (K. Blin *et al.*, 2013). We compared all the putative sub-clusters from our analysis against the known sub-clusters. To assess whether we identified a known sub-cluster, we calculated the fraction of the known sub-cluster that we captured in a putative sub-cluster as the overlap fraction. Setting the overlap fraction at 0.6, the sub-cluster motifs from the main dataset performs the best, identifying 77 (71%) of the validated sub-clusters, while the statistical sub-clusters from the main dataset and the sub-cluster motifs from the small dataset both identified 74 (Figure S11). The statistical sub-clusters of the small dataset had the worst performance capturing 63 validated sub-clusters.

Exploring the sub-cluster motifs

To showcase our findings we annotated 50 sub-cluster motifs, of which 23 originated from the set of known sub-clusters and 27 were annotated using MiBIG BGCs (Table S3). These annotations constitute 35 different substructures at different levels of detail (Figure 3). All 50 annotated sub-cluster motifs corresponded to SSCs to a certain degree, corroborating the sub-cluster motif annotations. Many of the annotated sub-cluster motifs are present in diverse BGC classes, while others occur in one class primarily (Figure S12;Figure S13). One example of the former is BGC0001597 (fluvirucin b2) that contains sub-cluster_motif_773 for a 3-amino-2-methylpropionyl starter unit constituting a macrolactam ring (Figure 3). This sub-cluster motif primarily occurs in NRPSs and type I PKSs. Interestingly, it also occurs in some Other class BGCs which cannot be classified by antiSMASH like NZ_KB913032.1.cluster021 and NZ_AXAS01000001.cluster006. This does not only provide these interesting BGCs with previously unknown structural information, it also adds to their validity.

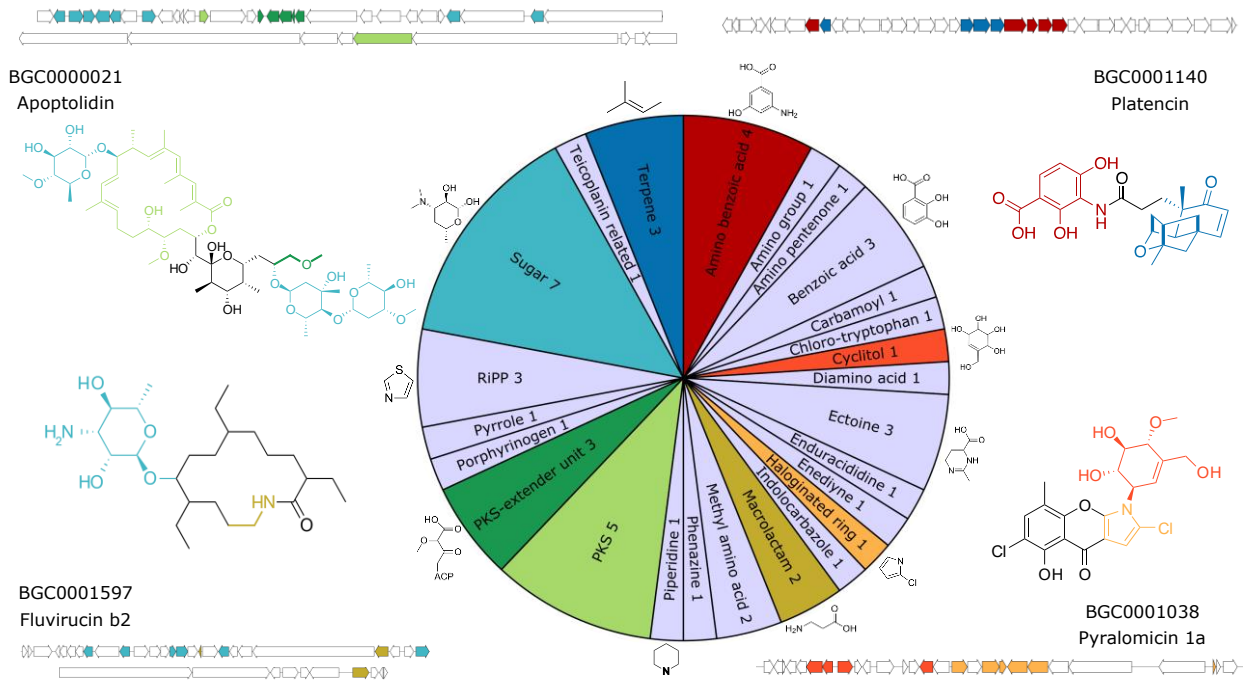

Figure 3 The pie chart visualises the annotations for the 50 sub-cluster motifs divided into general substructure groups, where an example substructure is shown for several groups. Additionally, examples of eight of the substructures are shown in the structures of apoptolidin, platencin, fluvirucin b2 and pyralomicin 1a, where the colour of the substructures correspond to the pie chart. For these four compounds, their respective BGCs are shown where the sub-cluster motifs are highlighted in the same colour as the substructures they encode.

Identifying the BGC for heronapyrrole biosynthesis

Information about the sub-clusters present in a BGC is not only useful to predict the product of a BGC, it could also be used as a tool to identify a BGC for a known compound. As an example, we aimed to identify the BGC responsible for heronapyrrole biosynthesis as the producing organism is present in the antiSMASH database and a candidate BGC has yet to be identified. Heronapyrroles A-D are a group of farnesylated nitropyrroles recently isolated from *Streptomyces* sp. CMB-StM0423 (Raju *et al.*, 2010; Schmidt *et al.*, 2014). As the heronapyrroles consist of a terpene-and a pyrrole derived moiety, we considered sub-cluster motifs related to terpenes and pyrroles. Based on antiSMASH classes, seven BGCs of CMB-StM0423 contain terpene moieties, of which one also contained a sub-cluster motif which we annotated as terpene related. One of these BGCs classified as a terpene, NZ_CP025407.1.cluster026, contained sub-cluster motif 972 as well as SCC_1465, which we annotated as pyrrole related based on the pyrrole moieties in kosinostatin and showdomycin (Ma et al., 2013; Palmu et al., 2017). NZ_CP025407.1.cluster026 is the only BGC in CMB-StM0423 that contains sub-cluster motif 972, which leads us to the hypothesis that it is the BGC responsible for heronapyrrole biosynthesis (Figure 4). This hypothesis is substantiated by the presence of the terpene elements that could be responsible for the farnesyl moiety. Additionally, we also identified sub-cluster_motif_972 and SCC_1465 in NZ_CP011492.1.cluster001. This predicted BGC originates from *Streptomyces* sp. CNQ-509, which is the producing strain of a group of farnesylated nitropyrroles called nitropyrrolins A-E, which are very similar to the heronapyrroles.

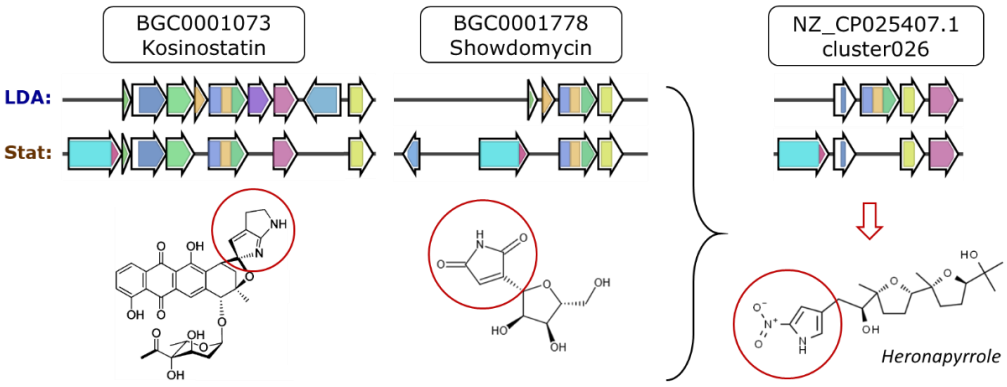

Figure 4 Sub-clusters from kosinostatin and showdomycin responsible for the biosynthesis of their pyrrole derivatives (Ma et al., 2013; Palmu et al., 2017). The lane LDA shows sub-clusters from sub-cluster motif 972, while the lane Stat shows sub-clusters from SCC 1465. On the right the hypothesis is depicted that NZ_CP025407.1.cluster026 is responsible for heronapyrrole synthesis based on the presence of the same sub-cluster motif and SCC as kosinostatin and showdomycin.

Correlation analysis

We deemed it interesting to assess if we could correlate substructures to sub-clusters in an automated manner as this could have the potential to link unknown molecules to BGCs at a large scale. We used a previously defined correlation score which assumes that a BGC is needed to synthesise a product, but that a BGC can be cryptic and not synthesise anything. Ernst *et al.* (2019) used the MS2LDA tool to create substructure models, called mass2motifs, from MS data of the Crüsemann dataset. For Crüsemann sub-cluster information, all Crüsemann BGCs were queried on the sub-cluster motifs and SCCs of the main dataset. For each of the 107,590 pairs of mass2motif and sub-cluster motif we calculated how well they co-occur across the Crüsemann strains with the correlation score, while we did the same for the 122,404 pairs of mass2motifs and SCCs. In order to prioritise interesting substructure-sub-cluster pairs, we performed permutation tests for all pairs. This resulted in 3,230 and 1,939 positive scoring combinations with a p-value below 0.1 for the mass2motif paired with sub-cluster motifs or SCCs, respectively (Figure 5). We identified 5 high correlation scores with low p-values between two staurosporine-related mass2motifs and both sub-cluster motifs and SCCs constituting the amino-sugar moiety of staurosporine. These are the only scores which we could identify as meaningful, which is already a good result as only a fraction of the mass2motifs, sub-cluster motifs and SCCs are annotated.

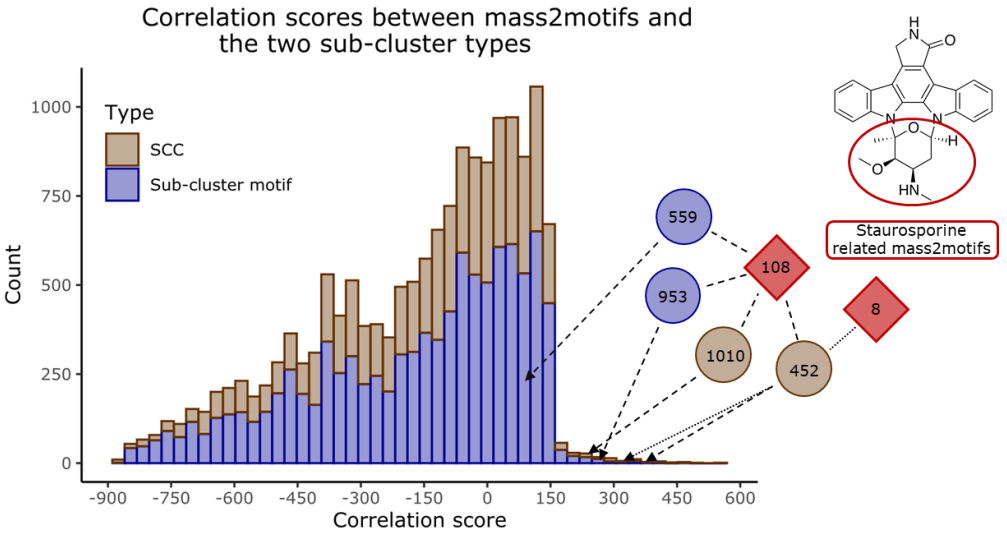

Figure 5 Stacked histogram of the correlation scores across the Crüsemann strains between the mass2motifs paired with either the SCCs or sub-cluster motifs with a p-value below 0.1. Highlighted with their scores are the pairs mass2motif_108 with SSC_452, SSC_1010, sub-cluster_motif_953 and sub-cluster_motif_559, and the pair

mass2motif_8 with SSC_452. The aforementioned sub-cluster motifs (blue) and SCCs (brown) are responsible for sugar synthesis in staurosporine, while both mass2motifs (red) are staurosporine related.

Discussion

The aim of this project was to improve upon the previous method for sub-cluster detection. To enhance BGC processing prior to sub-cluster detection, we used (sub)Pfam domains to represent sequence similarity which increased the scalability. Additionally, we reduced phylogenetic bias by filtering out redundant BGCs. By grouping the statistical sub-clusters in SCCs and removing redundant sub-clusters, we partly resolved the nested sub-cluster structures that result from the previous statistical method. Furthermore, we used LDA as a novel tool for sub-cluster detection. The sub-cluster motifs found with LDA had the highest benchmarking score on the characterised sub-clusters from SubClusterBlast. Moreover, we could annotate 50 sub-cluster motifs with substructure information, showing that LDA is a valuable method for the detection of sub-clusters.

Comparing LDA to the statistical method, they share a common goal as they both aim to find groups of co-occurring genes. LDA learns distributions over the domain-combinations from the data, which constitute sub-cluster motifs that provide a nicely clustered structure of similar sub-clusters, hereby capturing sub-cluster variation. The statistical method, however, creates many combinations of co-occurring genes, producing individual sub-clusters that exhibit highly nested structures making it harder to find similar sub-clusters across BGCs. Apart from the nested structures, the statistical method produces a huge amount of sub-clusters of which only a fraction probably provides meaningful information. This is illustrated by the fact that the statistical sub-clusters are very noisy. In a group of BGCs sharing multiple sub-clusters, all combinations of these shared sub-clusters would form new sub-clusters, which happens frequently. However, LDA generates a more limited amount of sub-cluster data, which might contain less meaningful sub-clusters compared to statistical method in absolute numbers, but has a way higher ratio of valid sub-cluster information. We partly solved the problems for the statistical sub-clusters by grouping them into SCCs and removing redundant sub-clusters, but problems still exist for the statistical method. Compared to LDA, it is for example rather difficult to query a BGC using the statistical sub-clusters. This is partly due to the fact that it would quickly become very time consuming to query for statistical sub-clusters while allowing inexact matching. For these reasons we now propose to use LDA as the main method for sub-cluster detection as it captures sub-cluster variety in the sub-cluster motifs and can be used easily to query BGCs for sub-cluster motifs. The statistical sub-clusters could still be used to identify the sub-cluster boundaries better, by for example clustering them within the sub-cluster motifs. In doing so, when a BGC matches a certain sub-cluster motif, it could be checked if that BGC contains any of the statistical sub-clusters clustered within the certain sub-cluster motif, hereby improving accuracy of the sub-cluster detection. The drawback of the statistical method that it produces highly nested and variable sub-clusters could as such be used as a strength.

The fact that a sub-cluster motif is a distribution over genes allows for a fast way to query BGCs for the presence of sub-cluster motifs. This also highlights sub-cluster motifs from a biological point of view. Sub-cluster motifs contain a few main genes responsible for the core of a substructure and have the ability to capture many genes that modify and diversify this core. Such is the case for sugar motifs like sub-cluster motif 842, where most sugars constitute dideoxy-sugars that are sometimes aminated or methyl-aminated. However, the sugar sub-cluster motifs also provide an example for a downside of the unsupervised LDA method. Although there are some structural differences between different sub-cluster motifs for sugars, the environment of the sugar sub-clusters had an impact on shaping the sub-cluster motifs. The sugar-related sub-cluster motif 72 contains for example a number of type II PKS genes, while the sugar-related sub-cluster motif 743 contains diazo-group genes like in lomaivicitin. A way to solve this would be to apply LDA in a semi-supervised manner, which is a huge asset of LDA. Before training an LDA model, certain motifs could be seeded beforehand, which allows accurate sub-cluster motifs to be reused in new analyses, analogous to MotifDB, where annotated mass2motifs are stored (Rogers *et al.*, 2019). Such semi-supervised approaches would allow for noise to be eliminated from sub-cluster motifs and sub-cluster motifs to be finetuned.

Another way to reduce noise and to identify the more robust sub-cluster motifs would be to train multiple LDA models. Sub-cluster motifs that are found in every LDA model would constitute conclusive sub-cluster motifs, whereas sub-cluster motifs that are identified a majority of the time would still be considered reasonably accurate. In this manner, noisy sub-cluster motifs that arise through chance would be filtered out, as they would only occur in one of the many LDA models. Noisy genes in accurate sub-cluster motifs could be filtered out by taking intersects of multiple similar sub-cluster motifs. As another option, each BGC could be represented multiple times in training to increase the observations of less frequently occurring sub-clusters. This could lead to better estimation of the sub-cluster motif distributions over the

data and cause less erroneous mixed sub-cluster motifs. We tried this shortly for the small dataset and saw that the overlap with SubClusterBlast increased slightly. This would be interesting to continue experimenting with in the future.

The 50 sub-cluster motifs that we annotated could readily be integrated into tools like antiSMASH to enhance structure predictions. They could also prove highly useful to classify BGCs for which there is currently no class annotation. In the case of antiSMASH, including these 50 sub-cluster motifs would mean a vast improvement to the current scope of sub-cluster prediction, as 27 out of the 50 sub-cluster motifs were not included in the list of SubClusterBlast sub-clusters. Additionally, the 23 sub-cluster motifs that showed overlap with the SubClusterBlast sub-clusters could be used as a replacement for the SubClusterBlast method. As querying a BGC for sub-cluster motifs is rather fast, this could result in a substantial speed increase, which would have to be investigated in the future.

With our correlation analysis, we demonstrated that it is already possible to connect substructures with sub-clusters in an automated manner. However, the previously defined correlation analysis was not ideal for our situation. One of the problems was the limited amount of data, as we could only use 137 strains from the Crüsemann dataset, 50 annotated sub-cluster motifs and 40 annotated mass2motifs. Not only did we have a limited set of strains, all strains were highly related to each other, meaning that many compounds and BGCs are shared between them. By default this created high correlation scores for pairs of sub-clusters and substructures occurring in abundant BGCs and compounds. We aimed to solve this by performing permutations tests to assess the likelihood of a high scoring pair arising by chance, which is the case for very abundant pairs. This only left very few high scoring pairs, in which we could only identify the pairs related to staurosporine. Doroghazi *et al.* (2014) developed the correlation metric to allow for the fact that many BGCs can be cryptic by not punishing the absence of a structure when a BGC is present. Because of the nature of the scoring metric many pairs with low p-values and scores just above zero arose, which consist mainly of mass2motifs with very low degrees paired with all sub-cluster occurring in the same strains. Furthermore, this correlation method generally results in a lot of noise, as sub-clusters and substructures that occur in a shared subset of strains will all correlate to each other. Such co-correlating structures make the identification of the actual correlating pair therefore difficult, especially with limited annotations. Identifying clusters of co-correlating pairs could therefore provide a way to make the interpretation of this analysis easier. Additionally, the correlation analysis is not perfect in our case, as multiple different sub-clusters are often responsible for synthesising the same substructure. As an example, mass2motif 119 is annotated as a dimethyl-amino-deoxysugar found in both rosamicin and lomaivicitin. However, the sub-cluster responsible for the sugar group in rosamicin is present in a different sugar sub-cluster motif than the sub-cluster from lomaivicitin. This illustrates a big drawback for using this method. In order to solve this, sub-cluster motifs that constitute similar structures should be grouped together before running the correlation analysis. Combining this with the integration of more diverse species would improve this correlation analysis drastically. As this is just a first step in linking substructure- and sub-cluster models with limited information, we expect that analyses like these will have great impact in the future facilitating metabologenomics experiments.

Throughout our sub-cluster detection analyses, we did not find many sub-clusters containing multiple multi-domain genes such as PKS or NRPS modules. This is due to the fact that we tokenised each gene as a combination of domains, which does not allow for capturing small variations in multi-domain genes. To model such multi-domain genes it would probably be better to tokenise each BGC as a string of domains, ignoring gene boundaries, as is done by (Navarro-Muñoz *et al.*, 2018). Another generic issue in our analysis is that we did not include some important biosynthetic domains, which caused some sub-clusters not to be detected properly. An example is the sub-cluster for the indolocarbazole moiety in staurosporine, which was not detected because the main gene of this sub-cluster contains the Ferritin-like domain which was not included in the analysis.

In this project, we have provided an improved approach for the detection of sub-clusters. We demonstrated that LDA is an effective tool for the discovery of new sub-clusters. Using MiBIG BGCs, we were able to annotate 50 sub-cluster motifs with structural information. These annotated motifs can now be used in future experiments and for the improvement of structural predictions in BGC prediction frameworks like antiSMASH. In antiSMASH, the annotated sub-cluster motifs could provide an addition to SubClusterBlast, or even serve as a replacement. By linking the heronapyrroles and nitropyrroles to their putative producing BGC, we illustrated a direct application of our work. Additionally, we provided the initial step for linking sub-clusters to substructures in a systematic way, which in the feature could lead to automated connection of BGCs to their NPs.

References

Alborzi, S. Z., Devignes, M.-D., & Ritchie, D. W. (2017). ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC Bioinformatics, 18*(1), 107. doi:10.1186/s12859-017-1519-x

Arthur, D., & Vassilvitskii, S. (2007). *k-means++: The advantages of careful seeding.* Paper presented at the Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.

Bateman, A., Smart, A., Luciani, A., Salazar, G. A., Mistry, J., Richardson, L. J., . . . Hirsh, L. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research, 47*(D1), D427-D432. doi:10.1093/nar/gky995

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., & Weber, T. (2013). antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res, 41*(Web Server issue), W204-212. doi:10.1093/nar/gkt449

Blin, K., Pascal Andreu, V., de Los Santos, E. L. C., Del Carratore, F., Lee, S. Y., Medema, M. H., & Weber, T. (2019a). The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res, 47*(D1), D625-D630. doi:10.1093/nar/gky1060

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., . . . Weber, T. (2019b). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research, 47*(W1), W81-W87. doi:10.1093/nar/gkz310

Bron, C., & Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM, 16*(9), 575-577. doi:10.1145/362342.362367

Chen, X., Hu, X., Shen, X., & Rosen, G. (2010). *Probabilistic topic modeling for genomic data interpretation.* Paper presented at the 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

Chevrette, M. G., & Currie, C. R. (2018). Emerging evolutionary paradigms in antibiotic discovery. *J Ind Microbiol Biotechnol*. doi:10.1007/s10295-018-2085-6

Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., . . . Fischbach, M. A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell, 158*(2), 412-421. doi:10.1016/j.cell.2014.06.034

Crüsemann, M., O'Neill, E. C., Larson, C. B., Melnik, A. V., Floros, D. J., da Silva, R. R., . . . Moore, B. S. (2017). Prioritizing Natural Product Diversity in a Collection of 146 Bacterial Strains Based on Growth and Extraction Protocols. *J Nat Prod, 80*(3), 588-597. doi:10.1021/acs.jnatprod.6b00722

Dayan, F. E., Cantrell, C. L., & Duke, S. O. (2009). Natural products in crop protection. *Bioorganic & medicinal chemistry, 17*(12), 4022-4034.

Del Carratore, F., Zych, K., Cummings, M., Takano, E., Medema, M. H., & Breitling, R. (2019). Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. *Communications Biology, 2*(1). doi:10.1038/s42003-019-0333-6

Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K. S., Haines, R. R., Tchalukov, K. A., . . . Metcalf, W. W. (2014). A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol, 10*(11), 963-968. doi:10.1038/nchembio.1659

Ernst, M., Kang, K. B., Caraballo-Rodríguez, A. M., Nothias, L.-F., Wandy, J., Chen, C., . . . van der Hooft, J. J. J. (2019). MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites, 9*(7), 144.

Fischbach, M. A., Walsh, C. T., & Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. *Proceedings of the National Academy of Sciences, 105*(12), 4601. doi:10.1073/pnas.0709132105

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). *Online learning for latent dirichlet allocation.* Paper presented at the advances in neural information processing systems.

Katz, L., & Baltz, R. H. (2016). Natural product discovery: past, present, and future. *Journal of industrial microbiology & biotechnology, 43*(2-3), 155-176.

Kim, E., Moore, B. S., & Yoon, Y. J. (2015). Reinvigorating natural product combinatorial biosynthesis with synthetic biology. *Nat Chem Biol, 11*(9), 649-659. doi:10.1038/nchembio.1893

Li, J. W. H., & Vederas, J. C. (2009). Drug Discovery and Natural Products: End of an Era or an Endless Frontier? *Science, 325*(5937), 161. doi:10.1126/science.1168243

Ma, H.-M., Zhou, Q., Tang, Y.-M., Zhang, Z., Chen, Y.-S., He, H.-Y., . . . Tang, G.-L. (2013). Unconventional Origin and Hybrid System for Construction of Pyrrolopyrrole Moiety in Kosinostatin Biosynthesis. *Chemistry & Biology, 20*(6), 796-805. doi:https://doi.org/10.1016/j.chembiol.2013.04.013

Medema, M. H., Cimermancic, P., Sali, A., Takano, E., & Fischbach, M. A. (2014). A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput Biol, 10*(12), e1004016. doi:10.1371/journal.pcbi.1004016

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., . . . Glockner, F. O. (2015). Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol, 11*(9), 625-631. doi:10.1038/nchembio.1890

Menzella, H. G., Reid, R., Carney, J. R., Chandran, S. S., Reisinger, S. J., Patel, K. G., . . . Santi, D. V. (2005). Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. *Nature Biotechnology, 23*(9), 1171.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research, 41*(12), e121-e121.

Navarro-Muñoz, J., Selem-Mojica, N., Mullowney, M., Kautsar, S., Tryon, J., Parkinson, E., . . . Medema, M. H. (2018). A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. *bioRxiv*, 445270. doi:10.1101/445270

Palmu, K., Rosenqvist, P., Thapa, K., Ilina, Y., Siitonen, V., Baral, B., . . . Metsä-Ketelä, M. (2017). Discovery of the Showdomycin Gene Cluster from Streptomyces showdoensis ATCC 15227 Yields Insight into the Biosynthetic Logic of C-Nucleoside Antibiotics. *ACS Chemical Biology, 12*(6), 1472-1477. doi:10.1021/acschembio.7b00078
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine Learning research, 12*(Oct), 2825-2830.
Raju, R., Piggott, A. M., Barrientos Diaz, L. X., Khalil, Z., & Capon, R. J. (2010). Heronapyrroles A−C: Farnesylated 2-Nitropyrroles from an Australian Marine-Derived Streptomyces sp. *Organic Letters, 12*(22), 5158-5161. doi:10.1021/ol102162d
Rehurek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora.* Paper presented at the In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.
Rogers, S., Ong, C. W., Wandy, J., Ernst, M., Ridder, L., & van der Hooft, J. J. J. (2019). Deciphering complex metabolite mixtures by unsupervised and supervised substructure discovery and semi-automated annotation from MS/MS spectra. *Faraday Discussions, 218*(0), 284-302. doi:10.1039/C8FD00235E
Schmidt, J., Khalil, Z., Capon, R. J., & Stark, C. B. W. (2014). Heronapyrrole D: A case of co-inspiration of natural product biosynthesis, total synthesis and biodiscovery. *Beilstein Journal of Organic Chemistry, 10*, 1228-1232. doi:10.3762/bjoc.10.121
Skinnider, M. A., Merwin, N. J., Johnston, C. W., & Magarvey, N. A. (2017). PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res, 45*(W1), W49-W54. doi:10.1093/nar/gkx320
Traxler, M. F., & Kolter, R. (2015). Natural products in soil microbe interactions and evolution. *Natural product reports, 32*(7), 956-970.
van der Hooft, J. J., Wandy, J., Barrett, M. P., Burgess, K. E., & Rogers, S. (2016). Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci U S A, 113*(48), 13738-13743. doi:10.1073/pnas.1608041113
Zhang, M.-Q., Gaisser, S., Nur-E-Alam, M., Sheehan, L. S., Vousden, W. A., Gaitatzis, N., . . . Martin, C. J. (2008). Optimizing Natural Products by Biosynthetic Engineering: Discovery of Nonquinone Hsp90 Inhibitors. *Journal of Medicinal Chemistry, 51*(18), 5494-5497. doi:10.1021/jm8006068

Supplementary data

Supplementary methods

Tokenising BGCs

To represent sequence similarity, BGCs were tokenised by converting them into strings of Pfam domains, using the HMMER3 tool hmmscan and the Pfam database version 32.0 (Bateman *et al.*, 2018; Mistry *et al.*, 2013). As Pfams are very broad domain models, we divided the Pfams that are most important for BGCs into more specific domain models called 'subPfams', to increase the resolution for sub-cluster detection. To create subPfams, a Pfam is divided into more narrow domains models that cover the subspaces of that Pfam, by extracting the multiple sequence alignment of a Pfam and separating it into clades. A new profile Hidden Markov Model (pHMM) is then built for each clade, each of which constitutes a subPfam (Figure S6A). The 112 biosynthetic Pfams that are most abundant in the antiSMASH database were converted into subPfams (Kautsar, S. A., personal communication, Supplementary files). We created our own pHMM database by replacing these 112 Pfams with their corresponding subPfams in the Pfam database version 32.0. To query a BGC, we used hmmscan to scan against our pHMM database with the tc-cutoff as a cutoff on the bitscore. Multiple hits in a gene were allowed to overlap by 10%. If the overlap was higher only the hit with the highest bitscore was kept. In this fashion, we tokenised each BGC as a string of genes, where each gene is a token represented as a combination of the present domains (Figure S6B). Genes without a hit were represented by a dash.

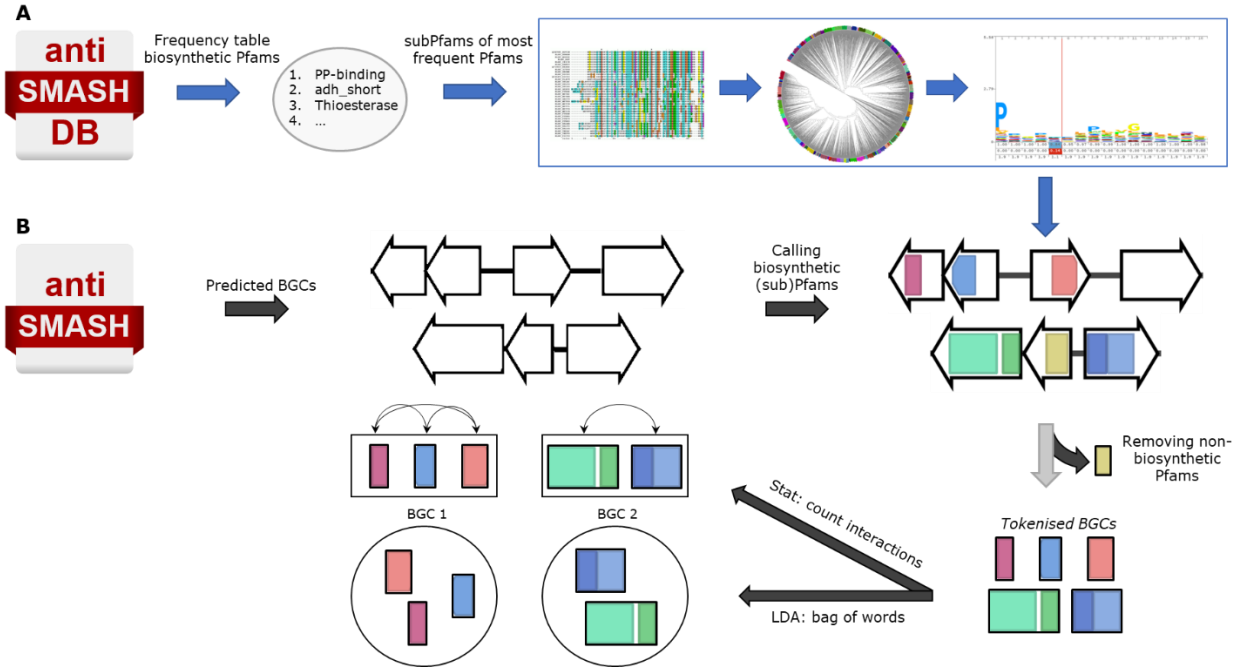

Figure S6 (A) subPfams are constructed for the 112 most frequent Pfam domains in the antiSMASH database by dividing the multiple sequence alignment of a Pfam into clades and converting each clade into a new pHMM. (B) The BGCs predicted by antiSMASH are tokenised by detecting (sub)Pfams in each gene, where non-biosynthetic Pfams are removed. After tokenising the BGCs, sub-cluster can be detected with the statistical method (Stat), where the tokenised genes are represented in their original order, or by LDA, which assumes a bag of words model where original gene order is not taken into account.

Filtering BGCs

In order to reduce phylogenetic bias, we filtered out redundant BGCs by constructing a similarity network of BGCs and choosing representative nodes from this network. As a similarity measure between BGCs we used an Adjacency Index of domains (AI), which has been used previously to assess BGC similarity (Navarro-Muñoz *et al.*, 2018). The AI between BGCs is calculated by dividing the number of all distinct shared pairs of adjacent domains by the total number of distinct pairs of adjacent domains, while ignoring gene boundaries. We constructed undirected graphs of similar BGCs by connecting two BGCs if their AI was above 0.95. We also connected two BGCs if one BGCs was fully contained in the other. To select representatives from the graphs all maximal cliques in the graph are found using find_cliques from the networkx module, which is based on the algorithm described by Bron *et al.* (1973). Then, the BGC with the most domains is chosen from each maximal clique to remain in the analysis, iterating over the cliques from largest to smallest until there are no cliques left. If there is more than one BGC to choose from, the BGC with least connections is picked to stay in the analysis to preserve as much information as possible. BGCs in a clique that are not selected are filtered out. This process is repeated until there are no connections left between BGCs.

Filtering domains

As we are interested in groups of genes that are directly responsible for the biosynthesis of chemical substructures, we chose to only detect sub-clusters of biosynthetic genes. In order to only select such genes, we discarded all Pfams that were not present in a list of 1,839 biosynthetic Pfams. We compiled this list by collecting all 3,010 EC-associated Pfams from ECDomainMiner using the lowest threshold (Alborzi *et al.*, 2017). We discarded domains from this list if they did not occur within existing pre-calculated BGCs (Kautsar, S. A., personal communication). This list was filtered further by searching for keywords like transporter or DNA-binding. We then added 50 manually curated biosynthetic domains to the list that were not part of ECDomainMiner but were frequent in the antiSMASH database, resulting in the list of 1,839 biosynthetic domains (Supplementary files). Additionally, Pfams were removed before sub-cluster detection if they occurred less than three times throughout the dataset. Subsequently, we removed all BGCs that contained less than two non-empty genes as result of Pfam filtering.

Clustering statistical sub-clusters

As the statistical method results in a large number of sub-clusters, we clustered them into sub-cluster families (SCFs) and the SCFs into sub-cluster clans (SCCs). To do so, we used the K-means algorithm implemented in scikit learn with k-means++ seeding, as it is very fast and easy to use on our large dataset of sub-clusters (Arthur *et al.*, 2007; Pedregosa *et al.*, 2011). We represented all sub-clusters as a presence/absence matrix with ones and zeros on which we ran K-means with 1,000 iterations and 20 restarts. For the construction of SCFs, we assessed the K-means clustering of different numbers for k. We chose a clustering based on the lowest within cluster sum-of-squares (WCSS), while keeping the amount of families to a minimum and trying to make sure one big 'hairball' cluster is formed with unrelated sub-clusters. In order to cluster the SCFs into SCCs, we clustered the centroids from the SCF clustering and assessed the clustering of different numbers for k in the same way as for the SCFs. We deemed an SCF to be meaningful if it had three genes that were present in at least 60% of the sub-clusters in the SCF. Additionally, we removed redundant sub-clusters from each SCF. We deemed a sub-cluster redundant if it had the same occurrence as a bigger sub-cluster in which it was contained completely.

Benchmarking against SubClusterBlast

The 127 SubClusterBlast sub-clusters were extracted from https://bitbucket.org/antismash/antismash/src/master/antismash/generic_modules/subclusterblast/subclusters.txt (K. Blin *et al.*, 2013). From the 127 validated sub-clusters, 109 had matching accessions in the MiBIG database. To see how many known sub-clusters we could identify, we calculated the overlap between all known sub-clusters and the putative sub-clusters from one of the detection methods. We defined an overlap as the number of genes (domain combinations) from a known sub-cluster that are present in a putative sub-cluster, divided by the number of genes in the known sub-cluster. We considered a known sub-cluster to be detected if there was at least one putative sub-cluster matching the known sub-cluster with an overlap above 0.6.

Annotation

The annotation of sub-cluster motifs or sub-cluster clans (SCCs) with substructures is still a manual task with low throughput, which is why we annotated only a small number of sub-cluster motifs and SCCs. In order to assign a substructure to a sub-cluster motif or SCC, we looked at the sub-cluster motifs and SCCs present in MiBIG BGCs, as their structures are validated. We considered an annotation appropriate for a sub-cluster motif or SCC if it is present in multiple MiBIG BGCs that share a similar substructure, while the genes in the sub-cluster comply with their proposed function in literature (Supplementary files). The latter is more valid for sub-cluster motifs and SCCs encompassing known sub-clusters as the genes from known sub-clusters are experimentally validated. To visualise and inspect identified sub-clusters, we improved an existing BGC visualisation script from Navarro-Muñoz *et al.* (2018) for the purpose of sub-cluster visualisation.

Correlation analysis

In order to correlate substructures to sub-clusters in a systematic manner, we used the Crüsemann dataset to link substructure models to the two different sub-cluster models derived in this research, using a previously defined correlation metric (Doroghazi *et al.*, 2014). The substructure models constitute 300 mass2motifs generated previously with the MS2LDA tool, based on MS/MS data from the Crüsemann dataset (Ernst *et al.*, 2019). The two sub-cluster models were generated by querying all tokenised Crüsemann BGCs to the LDA model trained on the whole dataset, and to the SCCs generated from the whole dataset, respectively. A Boolean vector was created for each mass2motif, sub-cluster motif and SCC, representing the presence/absence in all strains of the Crüsemann dataset. We excluded motifs or clans if they were present in less than two strains. Each pair of mass2motif and sub-cluster motif or SCC was scored for a mutual presence/absence pattern across strains. This correlation score constitutes scoring +10 if both members of a pair are present in a strain, +1 if both members of a pair are absent in a strain, -10 if the mass2motif is present in a strain while a sub-cluster motif or SCC is not, or 0 if the mass2motif is absent in a strain while a sub-cluster motif or SCC is present. We prioritised valuable pairs by assessing how meaningful a positive score is in two ways: calculating the maximum possible correlation score without changing the occurrences, and performing a permutation test. The permutation test was carried out by

scrambling each Boolean vector 10,000 times, calculating 10,000 random scores for each pair and dividing the times a higher or equal score than the observed score occurs by 10,000.

Figures

Figure S7 Graphical representation of graph-based filtering for the small dataset: MiBIG-and Crüsemann BGCs. Each node represents a BGC and an edge represents an AI of 0.95 or higher. In blue are the BGCs chosen as representatives, while BGCs that are filtered out are in black.

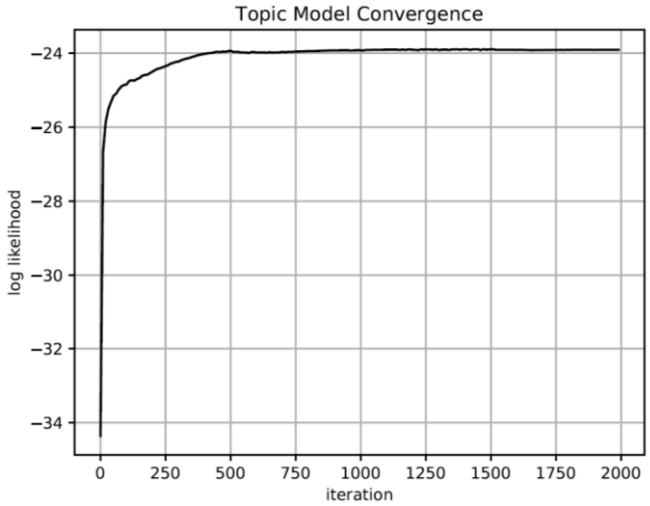

Figure S8 Convergence of the log-likelihood of the LDA model trained on the filtered 60,028 BGCs from the antiSMASH database, the Crüsemann dataset and the MiBIG database with 2,000 iterations of chunksize 3,000. Log-likelihood based on 28 held out BGCs.

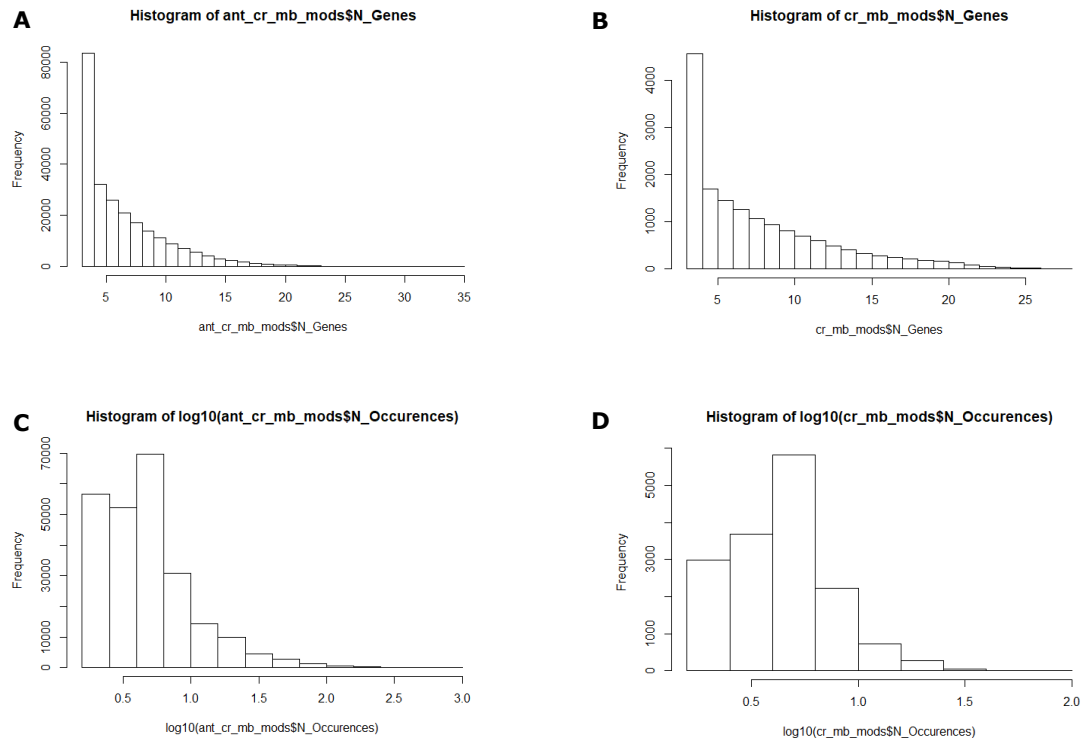

Figure S9 The distribution of the number of genes per module in the main dataset (A) and the small dataset (B), together with the distribution of the log10 of the occurrence in the main dataset (C) and the small dataset (D).

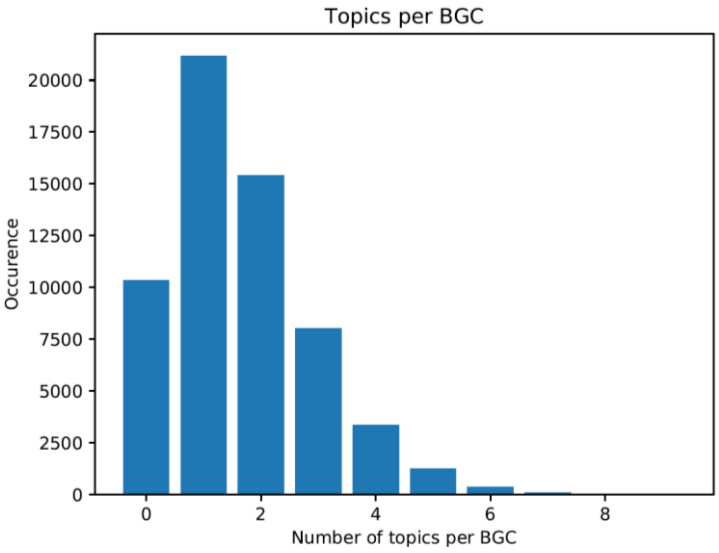

Figure S10 The number of topics or sub-cluster motifs per BGC in the main dataset, not counting sub-clusters of length one as these are almost definitely noise.

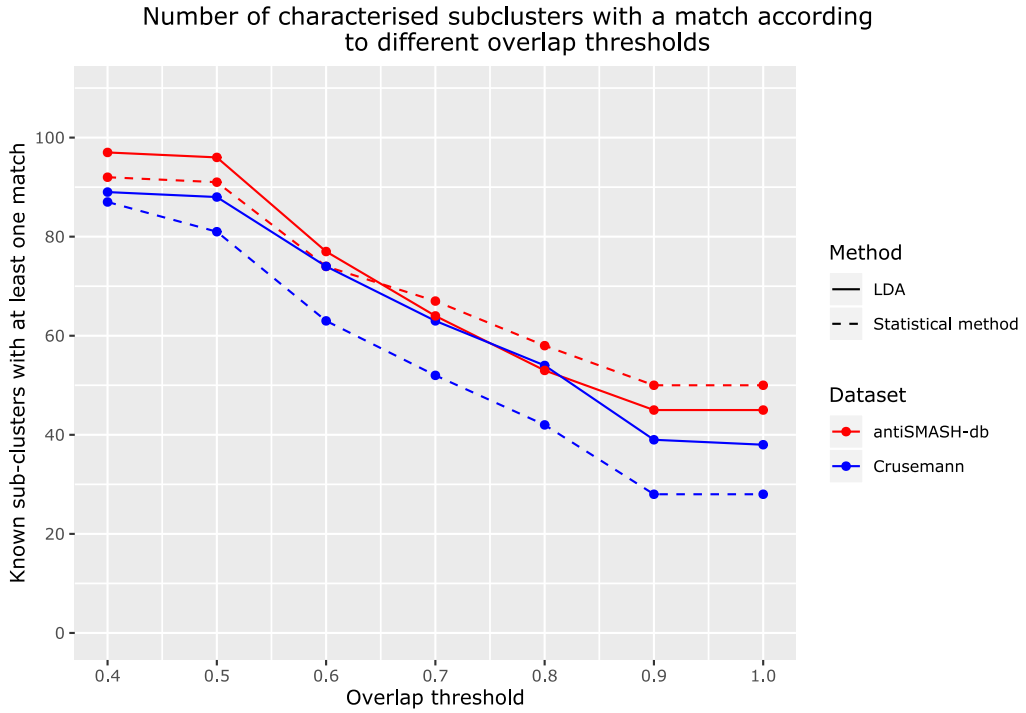

Figure S11 Overlap between SubClusterBlast and both sub-cluster detection methods applied on the main dataset (antiSMASH database) or the small dataset (Crüsemann), according to different overlap cut-offs. Both datasets also contain the MiBIG database.

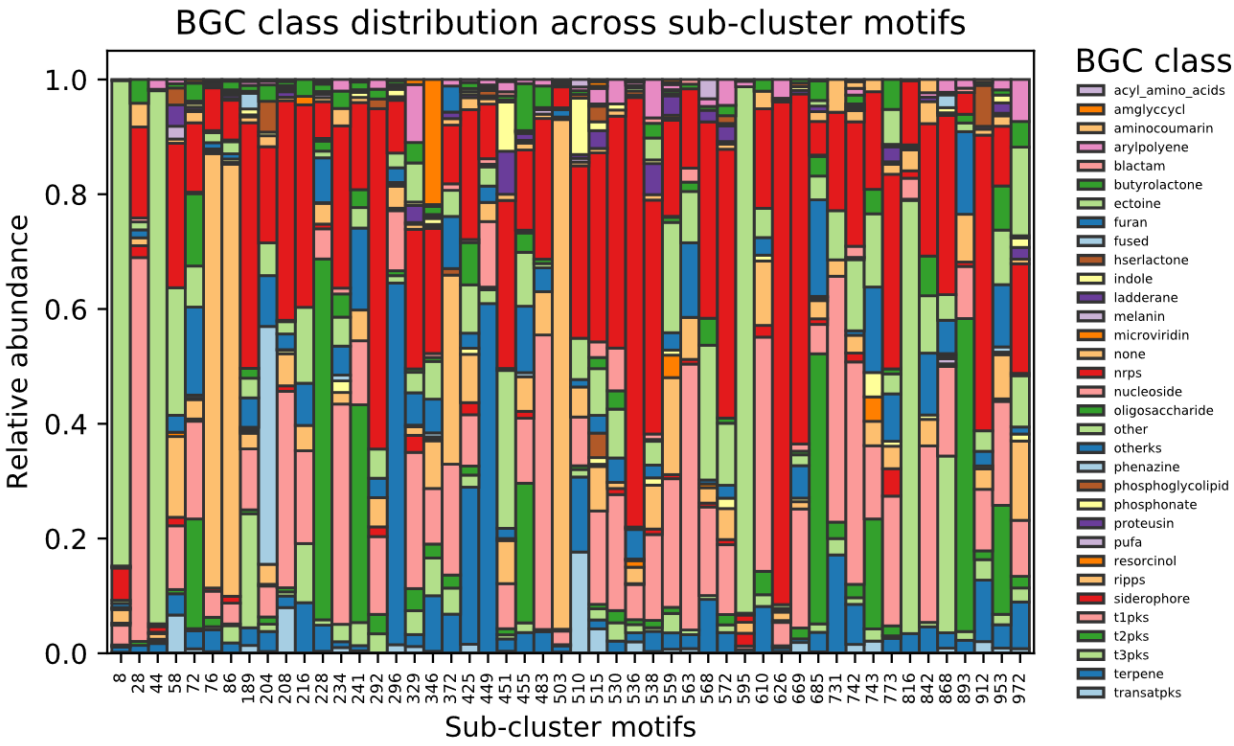

Figure S12 Relative abundance of antiSMASH classes when querying the main dataset (filtered) on the 50 annotated sub-cluster motifs. Matches of length 1 are ignored and hybrid class BGCs are counted for all classes they contain. Ripps classes are grouped together.

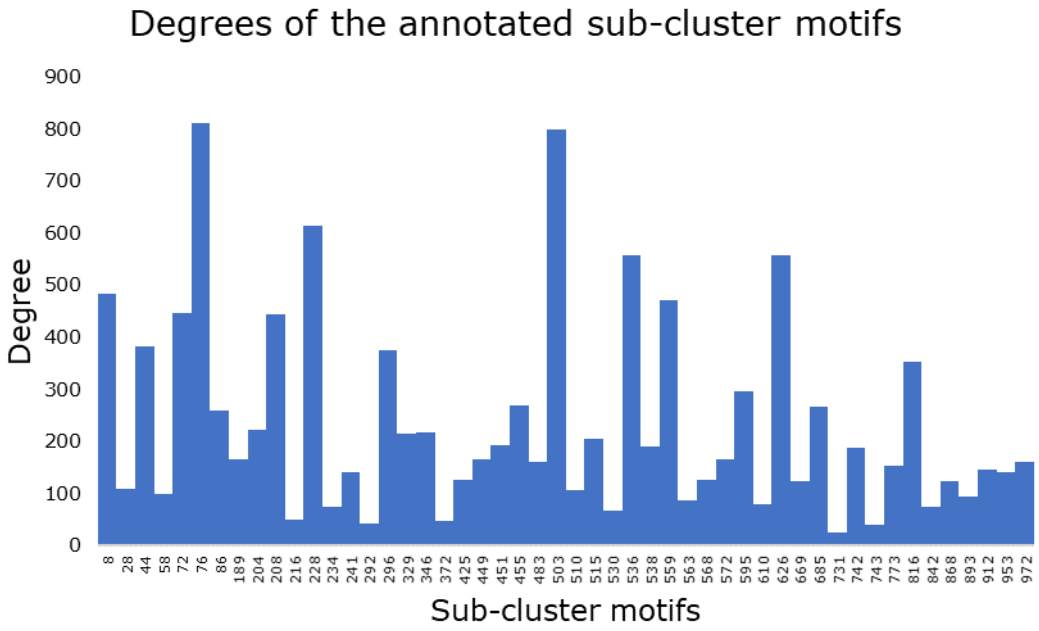

Figure S13 Degrees (occurrences) of the annotated sub-cluster motifs based on the main (filtered) dataset.

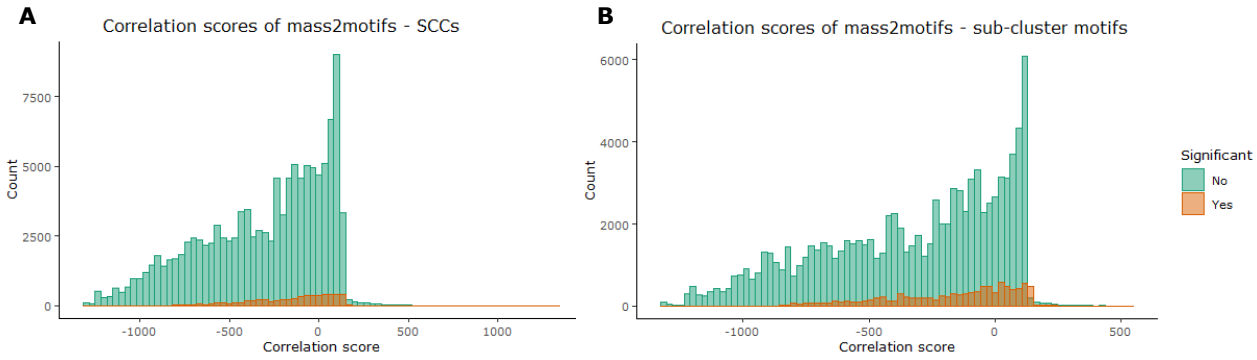

Figure S14 Correlation scores between mass2motifs and SCCs (A) or sub-cluster motifs (B), where the significant pairs are highlighted.

Tables

Table S1 Number of BGCs in the different datasets during different processing steps before sub-cluster detection. The main dataset is a combination of the antiSMASH database, the MiBIG database and the Crüsemann dataset. The small dataset combines the Crüsemann dataset with the MiBIG database.

Number of BGCs	antiSMASH-db	MiBIG	Crüsemann	Main dataset	Small dataset
Initial	152,122	1,819	5,927	159,868	7,746
On contig edge	41,914	0	1,367	43,281	1,367
Filtered	50,296	317	3,113	56,559	3,456
Final	59,912	1,502	1,447	60,028	2,923

Table S2 Equations for statistical method.

Equation 1 Hypergeometric equation for adjacency interactions between gene A and gene B. B_1: gene B not adjacent to gene A, B_2: gene B adjacent to gene A, B_3: gene B adjacent to gene A on both sides. N_1, N_2 and N_3 represent all available positions in these three categories, while N_{tot} represent all positions, and B_{tot} all occurrences of gene B.	$\boldsymbol{P_d = \frac{\binom{N_1}{B_1}\binom{N_2}{B_2}\binom{N_3}{B_3}}{\binom{N_{tot}}{B_{tot}}}}$
Equation 2 Hypergeometric equation for co-localisation interactions between gene A and gene B. B_1: gene B not co-localised with gene A, B_2: gene B co-localised to gene A, B_{nmax}: gene B co-localised with nmax gene A. N_1, N_2 and N_{nmax} represent all available positions in these three categories, while N_{tot} represent all positions, and B_{tot} all occurrences of gene B.	$P_d = \frac{\binom{N_1}{B_1}\binom{N_2}{B_2}\dots\binom{N_{n_{max}}}{B_{n_{max}}}}{\binom{N_{tot}}{B_{tot}}}$
Equation 3 Simplified hypergeometric equation for co-localisation interactions between gene A and gene B. B_1: gene B not co-localised with gene A, B_2: gene B co-localised to gene A. N_1 and N_2 represent all available positions in these two categories, while N_{tot} represent all positions, and B_{tot} all occurrences of gene B.	$P_d = \frac{\binom{N_1}{B_1}\binom{N_2}{B_2}}{\binom{N_{tot}}{B_{tot}}}$
Equation 4 Calculation of the p-value for an interaction. i: amount of interaction, i_{obs}: observed amount of interaction.	$p = P_{i \geq i_{obs}} = 1 - P_{i \leq i_{obs}} = 1 - \sum_{i \leq i_{obs}} P_d$

Table S3 Annotation table for the sub-cluster motifs. Detailed version in Subcluster_annotations.xlsb.

Sub-cluster motif	Annotation specific	Annotation general	Annotation grouping	Sub-Cluster -Blast	Degree	MiBIG evidence
8	ectoine	ectoine	Ectoine	No	483	BGC0000853, BGC0000854
28	t1pks	t1pks	PKS	No	108	BGC0001396, BGC0000047, BGC0001648, BGC0000035, BGC0001812, BGC0001658, BGC0000087, BGC0001199, BGC0000086, BGC0000052, BGC0000053, BGC0000144, BGC0001533, BGC0001662, BGC0001830, BGC0000123, BGC0000097, BGC0000038, BGC0000029, BGC0000059, BGC0000021
44	ectoine	ectoine	Ectoine	No	382	BGC0000855, BGC0000858, BGC0000852
58	AHBA/3-HAA	amino_benzoic_acid	Amino benzoic acids	Yes	98	BGC0000213, BGC0000187, BGC0000679, BGC0001140, BGC0001156, BGC0001295
72	(amino)sugar	sugar	Sugar	Yes	446	BGC0001595, BGC0000102,+80others
76	thiopeptide	RiPP	RiPP	No	811	BGC0001753,+20
86	lassopeptide	RiPP	RiPP	No	258	BGC0001655, BGC0001674, BGC0001781, BGC0001539, BGC0000579, BGC0000575, BGC0001673, BGC0000578, BGC0001552, BGC0001507, BGC0001645, BGC0001493, BGC0001548, BGC0001550, BGC0001549
189	DPG/HPG/BHT	teicoplanin/balhimycin_related	Teicoplanin related	Yes	165	BGC0000290, BGC0000440, BGC0000441, BGC0000311 +10
204	PCA/PDC	phenazine	Phenazine	No	221	BGC0001302, BGC0001080, BGC0000935
208	Methoxy-malonyl-ACP	methoxymalonyl-ACP	PKS-extender unit	No	444	BGC0000020, BGC0001511, BGC0000040, BGC0001034, BGC0000021
216	L-4-methylproline	L-4-methylproline	Methyl amino acid	Yes	50	BGC0000397
228	cyclic-t2pks	t2pks	PKS	No	613	BGC0000279, BGC0000256, BGC0000230, BGC0000200, BGC0000190, BGC0001062, BGC0001376,+10others
234	macrolactam	macrolactam	Macrolactam	No	74	BGC0000029, BGC0000097, BGC0001522, BGC0000078, BGC0001452
241	amide-ring/ring-oxidations/-methylation	t2pks-tailoring	PKS	No	140	BGC0000279, BGC0000256, BGC0000230, BGC0000200, BGC0000190, BGC0001062, BGC0001376,+10others
292	enduracididine	enduracididine	Enduracid-idine	Yes	41	BGC0000388, BGC0000341
296	cyclic_sesqui/-tetraterpene	terpene	Terpene	No	375	BGC0000651, BGC0000653, BGC0000674
329	chloro/bromo-phenyl/pyrrole	halogenated_aromatic_ring	Haloginated ring	No	215	BGC0000130, BGC0000131, BGC0001819, BGC0000111, BGC0001500, BGC0001172, BGC0000128, BGC0000374, BGC0000127, BGC0001038, BGC0001159
346	valienol/ valienone/ validone	cyclitol	Cyclitol	Yes	216	BGC0001038, BGC0000723, BGC0000722, BGC0000701

372	2-amino-3-hydroxy-cyclopent-2-enone	2-amino-3-hydroxycyclopent-2-enone	Amino pentenone	Yes	48	BGC0000052, BGC0000213, BGC0001298, BGC0000187
425	terpenoid	terpenoid	Terpene	Yes	126	BGC0000632
449	DMAPP/GPP	terpene	Terpene	Yes	164	BGC0000654, BGC0001664, BGC0000665, BGC0001126, BGC0000668, BGC0001594, BGC0001501, BGC0001612, BGC0000666, BGC0001595, BGC0001140, BGC0001156
451	Chlorinated-tryptophan/-indolocarbazole	chloro-tryptophan	Chloro-tryptophan	Yes	193	BGC0000809, BGC0000822, BGC0000823, BGC0001333, BGC0001335, BGC0001337
455	hydroxy/-methoxy-benzenes	benzene modification	Benzoic acids	Yes	268	BGC0000216, BGC0000261, BGC0000394, BGC0000421, BGC0000422, BGC0000202, BGC0001693, BGC0000236, BGC0000240, BGC0000241 + 10 others
483	enediyne	enediyne	Enediyne	Yes	159	BGC0000081, BGC0000112, BGC0000150, BGC0000965, BGC0001008, BGC0001397, BGC0001584
503	lanthionine	lantipeptide	RiPP	No	798	BGC0000507, BGC0000509,+30
510	indolocarbazole	indolocarbazole	Indolo-carbazole	No	107	BGC0000813, BGC0000814, BGC0001224, BGC0001223, BGC0001336
515	carbamoyl	carbamoyl	Carbamoyl	No	204	BGC0000090, BGC0000074, BGC0000834 +10
530	aminosugar	sugar	Sugar	Yes	66	BGC0000809, BGC0001522, BGC0000880
536	3-HAA/DHBA	hydroxy-benzoic_acid	Benzoic acids	No	557	BGC0001213, BGC0000368, BGC0001437
538	methyl-aspartate/methyl-asparagine	methyl-aspartate	Methyl amino acid	Yes	189	BGC0001770, BGC0000876, BGC0001448, BGC0000429, BGC0000167
559	sugar	sugar	Sugar	Yes	470	BGC0000025, BGC0000052,+40others
563	methoxymalonyl-ACP	methoxymalonyl-ACP	PKS-extender unit	Yes	87	BGC0000994, BGC0000065, BGC0000090
568	Uroporphyrin-ogen_III	uroporphyrin-ogen_III	Porphyrin-ogen	No	125	BGC0000906, BGC0000905
572	4-methyl-3-hydroxyanthranilic_acid	4MHT	Amino benzoic acids	Yes	165	BGC0000296, BGC0000428, BGC0000303, BGC0000409
595	ectoine	ectoine	Ectoine	No	296	BGC0000859, BGC0000857, BGC0000860, BGC0000856
610	piperideine-derivative	piperideine	Piperidine	No	78	BGC0001296, BGC0001433, BGC0001293
626	2,3-dihydroxybenzoic acid	DHBA	Benzoic acids	No	557	BGC0001185, BGC0000343, BGC0001502, BGC0000451, BGC0000454, BGC0000309, BGC0000945, BGC0001345, BGC0000401
669	methylated-sugar	sugar	Sugar	Yes	124	BGC0000148, BGC0000362, BGC0000363, BGC0000364, BGC0000365, BGC0000769, BGC0000875
685	t2pks	t2pks	PKS	Yes	265	BGC0000221, BGC0000227, BGC0000245, BGC0000225, BGC0000233, BGC0000269,+10
731	amino/guadinino	amino/guadinino	Amino group	No	25	BGC0000052, BGC0001662, BGC0001700
742	3-amino-5-hydroxy-benzoicacid	AHBA	Amino benzoic acids	Yes	187	BGC0000020, BGC0001511, BGC0000090, BGC0000106
743	Aminosugar	sugar	Sugar	No	40	BGC0000240, BGC0000241, BGC0001693
773	3-amino-2-methyl-propionyl-starter	macrolactam	Macrolactam	No	153	BGC0000167, BGC0001770, BGC0001597, BGC0001101, BGC0000202, BGC0001658
816	t3pks	t3pks	PKS	No	352	BGC0001647, BGC0000282
842	Desosamine/Aminosugar/4_6-dideoxysugar	sugar	Sugar	Yes	74	BGC0001830, BGC0000054, BGC0000055, BGC0000033, BGC0000102, BGC0001503, BGC0000078, BGC0001008, BGC0000035, BGC0000047, BGC0001396, BGC0000085, BGC0001812
868	dihydroxyphenylglycine/dihydroxybenzoicacid	dihydroxyphenylglycine	Amino benzoic acids	Yes	124	BGC0001233, BGC0001066, BGC0001148, BGC0001635, BGC0001819, BGC0001807
893	malonyl-CoA	malonyl-CoA	PKS-extender unit	No	94	BGC0000279, BGC0000216, BGC0000261
912	Diamino-butyricacid	DABA	Diamino acid	Yes	146	BGC0000950, BGC0000951, BGC0001807
953	aminosugar/methylated_sugar	sugar	Sugar	Yes	140	BGC0000193, BGC0001812, BGC0000096, BGC0001452, BGC0001522, BGC0000055, BGC0000019, BGC0000825, BGC0000826, BGC0001074, BGC0000199, BGC0000212, BGC0000216, BGC0000141
972	pyrrole/indole	pyrrole/indole	Pyrrole	No	161	BGC0001073, BGC0001778, BGC0001595, BGC0000668, BGC0000824

Table S4 Correlation scores between mass2motifs and sub-cluster types for pairs present in staurosporine.

Mass2motif	Sub-cluster type	Score	Max score	% of max score	p-value
mass2motif_108	sub-custer_motif_953	249	438	0.57	0.000
mass2motif_108	sub-custer_motif_559	68	278	0.24	0.000
mass2motif_8	SCC_452	300	615	0.49	0.010
mass2motif_108	SCC_452	355	607	0.58	0.010
mass2motif_108	SCC_1010	215	572	0.38	0.010

Supplementary files

- biosynthetic_pfams.txt
- subPfams.txt
- subcluster_annotations.xlsb