Bioinformatics, 35(19), 2019, 3584–3591
doi: 10.1093/bioinformatics/btz127
Advance Access Publication Date: 20 February 2019
Original Paper

OXFORD

Sequence analysis

Inter-Modular Linkers play a crucial role in governing the biosynthesis of non-ribosomal peptides

Sherif Farag [1,2,*], Rachel M. Bleich[2], Elizabeth A. Shank[3,4], Olexandr Isayev[2], Albert A. Bowers[2] and Alexander Tropsha[2,*]

[1]Curriculum in Bioinformatics and Computational Biology, [2]Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, [3]Department of Biology and [4]Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on October 13, 2018; revised on February 12, 2019; editorial decision on February 13, 2019; accepted on February 17, 2019

Abstract

Motivation: Non-ribosomal peptide synthetases (NRPSs) are modular enzymatic machines that catalyze the ribosome-independent production of structurally complex small peptides, many of which have important clinical applications as antibiotics, antifungals and anti-cancer agents. Several groups have tried to expand natural product diversity by intermixing different NRPS modules to create synthetic peptides. This approach has not been as successful as anticipated, suggesting that these modules are not fully interchangeable.

Results: We explored whether Inter-Modular Linkers (IMLs) impact the ability of NRPS modules to communicate during the synthesis of NRPs. We developed a parser to extract 39 804 IMLs from both well annotated and putative NRPS biosynthetic gene clusters from 39 232 bacterial genomes and established the first IMLs database. We analyzed these IMLs and identified a striking relationship between IMLs and the amino acid substrates of their adjacent modules. More than 92% of the identified IMLs connect modules that activate a particular pair of substrates, suggesting that significant specificity is embedded within these sequences. We therefore propose that incorporating the correct IML is critical when attempting combinatorial biosynthesis of novel NRPS.

Availability and implementation: The IMLs database as well as the NRPS-Parser have been made available on the web at https://nrps-linker.unc.edu. The entire source code of the project is hosted in GitHub repository (https://github.com/SWFarag/nrps-linker).

Contact: alex_tropsha@unc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As the threat of antibiotic resistance continues to rise and the number of available treatments continues to decline, the need to develop novel antibiotics is greater than ever. Non-ribosomal peptides (NRPs) are specialized metabolites produced by bacteria and fungi, many of which have clinical applications as antibiotics (e.g. daptomycin, vancomycin), anticancer agents (e.g. bleomycin) and immunosuppressants (e.g. cyclosporin). NRPs are synthesized by non-ribosomal peptide synthetases (NRPSs), which are exceptional mega-enzymes. Each NRPS protein consists of multiple modules, which consist of multiple catalytic domains that work together to assemble highly complex, bioactive secondary metabolites. These modules are joined together by linkers or strings of amino acids (Baltz, 2006; Felnagle *et al.*, 2008; Mootz *et al.*, 2000).

Combinatorial biosynthesis of novel NRPs has been a longstanding goal in chemical biology. Five major strategies have been employed so far: (i) exchanging entire NRPS genes across different biosynthetic gene clusters (BGCs) (Baltz *et al.*, 2006; Coëffet-Le Gal *et al.*, 2006; Nguyen *et al.*, 2006); (ii) exchanging modules (Nguyen *et al.*, 2006); (iii) exchanging domains (Calcott *et al.*, 2014); (iv) exchanging sub-domains (Crüsemann *et al.*, 2013); (v) using well-defined exchange units (XUs) and not modules as functional units (Bozhüyük *et al.*, 2018). Common across all of these strategies is that the adenylation domain (A-domain) is either swapped or edited in place. Since the A-domain is responsible for activating the substrate that will be incorporated into the final peptide product, swapping or modifying it will potentially lead to the synthesis of a different peptide. Moreover, a recent study has shown that in addition to their gate-keeping function, Condensation-domains (C-domains) also exhibit a module specificity-regulatory role, which helps even further diversification of NRPs and other natural peptides (Meyer *et al.*, 2016). Unfortunately, most of the NRP analogues derived using these strategies have resulted in either lower yield or no yield relative to the wild type (Calcott *et al.*, 2014; Stevens *et al.*, 2005; Winn *et al.*, 2016).

One possible reason for the generally poor performance of these strategies could be due to an incomplete understanding of the importance of linkers within NRPSs. There are two types of linkers within NRPS assembly lines: the regions between domains known as Inter-Domain Linkers (IDLs) (Bhaskara *et al.*, 2013) and the regions between modules known as Inter-Modular Linkers (IMLs). Studies have shown that IDLs can play an essential role in maintaining cooperative inter-domain interactions, as the composition and length of linkers affect protein stability, folding and domain-domain orientation (Gokhale and Khosla, 2000; Robinson and Sauer, 1998). These and other studies have provided mechanistic insights and biochemical evidence of the importance of linker regions in controlling NRPS domain conformation and emphasize the relevance of linkers to combinatorial biosynthesis outcomes.

Overall, IDLs have been more well-studied (Beer *et al.*, 2014; Doekel *et al.*, 2008; Reger *et al.*, 2007; Wu *et al.*, 2009; Yu *et al.*, 2013) than IMLs (Lott and Lee, 2017; Tarry *et al.*, 2017). When considering IMLs, the rule of thumb has been to keep them intact and not to remove, edit or swap them. The assumption is that interfering with these linkers would prevent module-module association and therefore diminish product yield (Winn *et al.*, 2016). This has led to a high level of uncertainty about the importance of IMLs, and no IML database currently exists to facilitate their analysis.

In this study, we endeavored to address these deficiencies by scanning 39 232 bacterial genomes for potential NRPS BGCs and implementing a NRPS-Parser to extract and analyze all potential IMLs across this database. Using these data, we have established the first public IMLs database and investigated whether there is a relationship between each IML and its adjacent A-domains. Our chief objective was to develop a better understanding of the role of IMLs in NRPSs in order to enable more efficient rational design of novel NRPs.

2 Materials and methods

2.1 Study design and dataset

Two major bacterial genome databases were used in this study: NCBI prokaryotic RefSeq genomes and ENA Ensembl bacterial genomes databases, comprising 70 844 and 41 610 bacterial genomes, respectively. In addition to that we also used the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) repository, which contains 408 NRP BGCs (Medema *et al.*, 2015). Due to the large amount of overlap between the two databases, 39 232 unique bacterial genomes were ultimately analyzed. We then downloaded the corresponding genomes (GenBank format) from the NCBI Genomes FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/) and ran antiSMASH 3.0 (Weber *et al.*, 2015), a tool that identifies and annotates specialized metabolite BGCs for the extraction of NRPS BGCs. We then applied our tool, NRPS-Parser, on all identified NRP clusters and extracted all possible IMLs. Next, we established the first IMLs database. We conducted a comprehensive analysis on all extracted IMLs in our database and investigated whether there is a relation between the IML and the activated substrates of adjacent A-domains (Fig. 1).

2.2 IML NRPS-Parser

After identifying all possible NRP BGCs, a parser dedicated to extracting IMLs within NRPSs was developed and implemented. The parser extracts linkers in the following pattern: 'A1-linker-A2' where A1 and A2 refer to the activated amino acid substrates of the A domains from module 1 and module 2, respectively (Supplementary Fig. S1). The linker is defined as the segment of amino acids connecting these two successive NRPSs modules. All domain borders have been identified by antiSMASH 3.0 using profile Hidden Markov Models (pHMMs), which are based on multiple sequence alignments of experimentally characterized signature proteins or protein domains (proteins, protein subtypes or protein domains that are each exclusively present in a certain type of biosynthetic gene clusters).

2.3 Web-server

All extracted IMLs are available on the following web-server (https://nrps-linker.unc.edu). The web-server has two major functionalities: (i) a NRPS-Parser that helps to extract IMLs from uploaded antiSMASH-predicted NRPS BGCs, with support for both antiSMASH 3.0 and 4.0 outputs (Weber *et al.*, 2015), and (ii) a filterable, searchable and exportable IML database comprised of the 39 804 IMLs extracted in this study. The tool is implemented using Python 2.7 and the Flask micro-web framework. The web-server is hosted by Carolina-cloudapps, a platform for developing and deploying web applications managed by the University of North Carolina at Chapel Hill.

3 Results

3.1 IML extraction

Our overall goal was to investigate whether there is a relationship between NRPS IMLs and their adjacent A-domains. To do so, we used the well-annotated NRPS clusters from the MIBiG repository. Furthermore, we applied antiSMASH 3.0 to predict all potential NRPS BGCs from 39 232 genomes (Supplementary File S1). We then extracted all possible IMLs from the antiSMASH-predicted NRPS BGCs using our NRPS-Parser, which led to the extraction of 39 804 IMLs (902 from MIBiG NRPS clusters and 38 902 from predicted NRPS BGCs) (Supplementary Files S2, S3). The IML NRPS-Parser extracts linkers in the pattern 'A1-linker-A2', where A1 and A2 refer to the activated amino acid substrates of the A domains from module 1 and module 2, respectively, and the linker is the string of amino acids joining these two successive NRPSs modules (Supplementary Fig. S1). After obtaining this collection of IMLs, we then pursued two main questions: (i) How specific are IMLs with regards to particular pairs of amino-acid-incorporating modules? (ii) How well conserved are IMLs within and across genera?

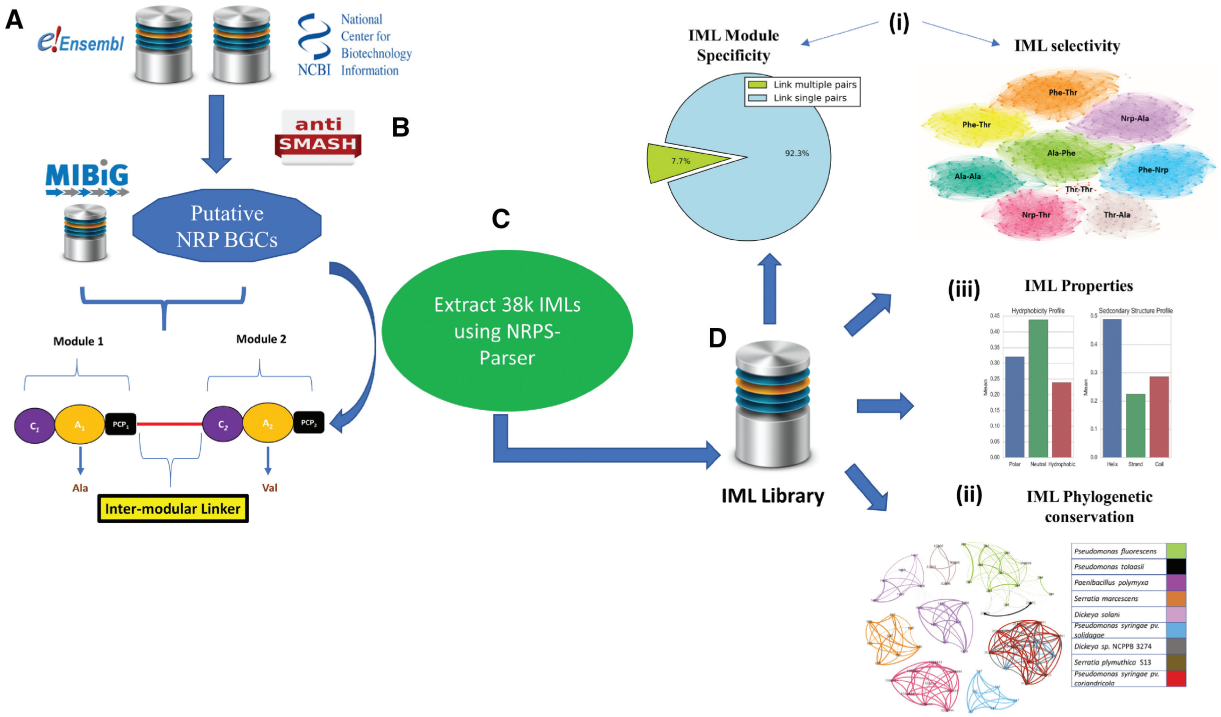

Fig. 1. Study design: (**A**) Two bacterial genome databases and a BCG repository were processed and integrated: The NCBI prokaryotic RefSeq genomes, ENA Ensembl Bacteria and MIBiG, respectively. (**B**) In addition to NRPS clusters from the MIBiG repository, antiSMASH 3.0 was run on downloaded genomes to identify all potential NRP BGCs. (**C**) Our NRPS-Parser tool was applied to extract all possible IMLs. (**D**) A database of IMLs was established. All identified IMLs were then analyzed for (i) Selectivity and specificity, (ii) Phylogenetic conservation and (iii) Properties

3.2 Analysis of IMLs

The 902 linkers obtained from the well-annotated MIBiG repository were extracted from 75 bacterial genera covering 196 species, while the 38 902 linkers extracted from the predicted NRPS BGCs were obtained from 138 bacterial genera covering 1956 bacterial species. When considering all of the extracted IMLs, their average GC nucleotide content was 13% (Supplementary Fig. S2) and their average length was 42 residues. For a deeper analysis of linkers length distribution, please refer to Figure 4 in the discussion section.

The amino acid characteristics of IMLs were composed, on average, of 44% neutral amino acids, 33% polar amino acids and 23% hydrophobic amino acids (Supplementary Fig. S3). This distribution agrees well with previous findings that linker regions tend to be less conserved in sequence and structure and contain more hydrophilic residues (Bae *et al.*, 2005; Udwary *et al.*, 2002). However, IMLs were found to exhibit more secondary structures than IDLs (Supplementary Fig. S3). A study by George and Heringa (2002) showed that the largest proportion of IDL residues, 38.3%, adopts the α-helical secondary structure, while 13.6% are in β-strands, 10.5% are in turns and the rest, 37.6%, are in coil or bend secondary structures. On the other hand, for IML residues 49% adopt the α-helical secondary structure, while 22% adopt the β-strands and the remaining 29% are found to be in coils. This finding demonstrates the difference between IDLs and IMLs while also reflecting their distinct functional roles in coordinating pairs of modules within NRPSs.

3.2.1 Selectivity of unique IMLs toward pairs of modules

In this analysis, pairs of modules are represented by their activated amino acid substrates. For example, Ser-Ala indicates that module 1 activates serine and module 2 activates alanine. Here, we investigated whether IMLs act as specific linkers (i.e. bridging particular pairs of modules) or as universal linkers with no specificity towards their modules. To do so, we first considered the number of unique module pairs to which a linker could bind. In order to avoid any ascertainment bias, we performed two preprocessing steps. First, we clustered all the extracted IMLs using clust-fast from UCLUST (Edgar, 2010). Next, we removed all singletons from the dataset, so as to investigate whether the same IML tends to bind the same pairs of modules. These preprocessing steps resulted in 3916 unique IML clusters. All clusters show less than 80% sequence similarity to each other. The pairwise centroid similarity distribution is depicted in Supplementary Figure S4.

Among all IML clusters, 92% (3616) were associated with only a single pair of modules (Supplementary Fig. S5A). For example, there are 427 occurrences of the linker 'SITDAAASQDDW VIVHDPE' in our database, which have been extracted from five different bacterial genera and are involved in the biosynthesis of 45 distinct NRPs. Each occurrence of this linker, regardless of genera or NRP product, links the same Gly-Cys module pair. Thus, a single IML typically bridges the same module pair. The remaining 8% of the linkers (300) tend to join only a limited number of module pairs (ranging between 2 and 13 unique pairs). For example, the linker 'ENTEVLPPIPLAPR', extracted from a single strain (*Burkholderia pseudomallei* 406e), bridges five distinct module pairs (Supplementary Fig. S5B). In addition, our analysis has shown that module pairs are not reversible: the IMLs between Ser-Ala modules differ from those that link Ala-Ser. Overall, it appears that IMLs are highly selective linkers in regard to the amino acids incorporated by their flanking modules.

An alternative way to illustrate the high level of IML selectivity is to examine the IMLs of a single bacterial species in a network. We selected *Mycobacterium abscessus* to illustrate this method, since it

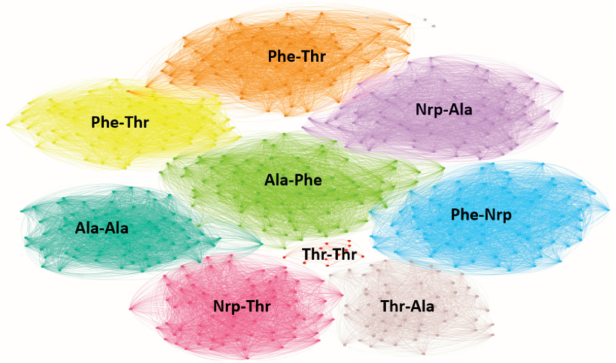

Fig. 2. Mycobacterium Abscessus Module-Specific IMLs: Community network, where nodes refer to linkers, and edges are constructed between two linkers, if they share 80% or more sequence similarity. The graph depicts nine distinct communities. Each community represents all the linkers that bind a specific pair of modules. For instance, the orange community refers to all linkers that bind the pair of modules that activate phenylalanine and tyrosine

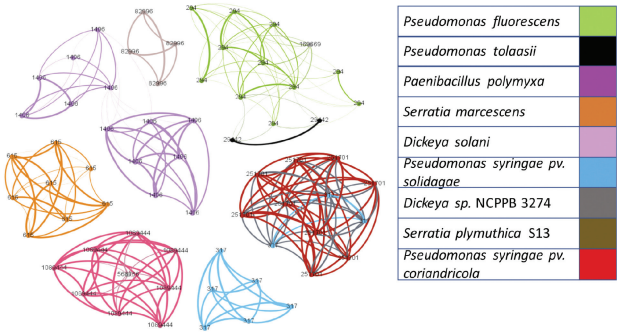

Fig. 3. (Thr-Val) IMLs community network: A network of all the linkers that bridge the Thr-Val module pair. These linkers belong to various distinct communities, despite the fact that they are all linking the same pair. The coloring of the graph refers to the species from which linkers were extracted

contains a reasonable number of linkers to depict in a two-dimensional network and its linkers bridged a range of distinct module pairs. In this network visualization, the IMLs are represented as nodes, with edges connecting nodes that show at least 80% similarity based on pairwise sequence alignment using the Needleman Wunch algorithm. We then applied a Louvain community detection algorithm (Blondel *et al.*, 2008), which detected nine distinct communities, each of which consisted of linkers that bind specifically a distinct pair of modules (Fig. 2, Supplementary Table S1). Similar results were obtained when we conducted the same analysis on linkers extracted from *Burkholderia pseudomallei* (Supplementary Fig. S6). These visualizations further supports the conclusion that module-specific IMLs dominate within NRPS BGCs.

3.2.2 Phylogenetic conservation of module-specific IMLs within and across genera

Our analysis so far indicates that IMLs are very selective towards module pairs. Here, we probe whether pairs of modules tend to be linked by the same IML regardless of the bacterial species from which they were extracted. We conducted an all-by-all comparison of module pairs vs. genera (computing the degree of conservation of IMLs linking a specific module pair both within and across genera) and then built a community network to visualize the phylogenetic distributions of IMLs that link the same module pair.

All-by-all comparison analysis: The NRPS BGCs from the MIBiG repository contain 116 unique module pairs. For every module pair we computed the similarity matrix of all its linkers using the Needleman Wunch algorithm, with an 80% similarity cut-off. Of the 2854 pairwise comparisons, just 10% (285) were able to reach or exceed the 80% similarity cut-off (Supplementary Fig. S7A). Of these 285 comparisons that were highly similar, 85% were from IMLs obtained from the same bacterial genus and 15% were from IMLs obtained from different genera (Supplementary Fig. S7A). Of the remaining 90% (2569) of comparisons that exhibited a low degree of conservation, 60% were between linkers extracted from different genera (Supplementary Fig. S7A). These findings indicate that module-specific IMLs tend to be more conserved within bacterial genera (Supplementary Fig. S7B), whereas multiple distinct IMLs exist that link the same module pair across different genera (Supplementary Fig. S7C). Furthermore, 83% of the IMLs that come from the same genera, yet show a low degree of conservation, were extracted from different species (Supplementary Fig. S7D). When we expanded this same analysis to the larger set of predicted NRPS BGCs, very similar results were obtained (Supplementary Fig. S8A). Both analyzed datasets show multiple cases of highly similar IMLs, if not completely identical, despite being extracted across distinct genera. The main reason behind such observation, is the horizontal gene transfer phenomena (HGT) which is the movement of genetic material between unicellular and/or multicellular organisms other than by the transmission of DNA from parent to offspring (vertical). For example, we found 27 instances of the IML 'VAL-ESKEEQTFEPIRQAP-ASP' across 3 different genera *Bacillus*, *Brevibacterium* and *Jeotgalibacillus* (Supplementary Fig. S8B). Another example revealed 427 instances of the IML 'GLY-SITDAAASQDDWVIVHDPE-CYS' across 4 different genera *Citrobacter*, *Escherichia*, *Klebsiella* and *Enterobacter* (Supplementary Fig. S8C).

Community network visualization of Thr-Val IMLs: We constructed a community network visualization to illustrate the phylogenetic specificity of IMLs. We took all IMLs for Thr-Val pair obtained across all species and created a graph as described above. After applying the Louvain community detection algorithm (Blondel *et al.*, 2008) to this data, the nodes were colored based on the bacterial species they were obtained from. If IMLs were globally conserved across many bacterial species, we would expect to obtain a single large community network with multi-colored nodes. If instead IMLs were conserved within a single bacterial species, we would expect multiple distinct communities to be detected, where nodes within each community would have the same color. The data indicate that the latter is the case, underscoring the phylogenetic specificity of IMLs (Fig. 3).

3.2.3 IMLs as independent building blocks

We next wanted to explore whether IMLs were highly associated with single NRP products, or whether the same IML was involved in the biosynthesis of distinct NRPs. If the latter was the case, then IMLs could potentially act as biosynthetic building blocks to generate novel NRPs. To begin this analysis, we first needed to define the NRP products produced by our extracted NRPS BGCs. The NRPs from the MIBiG repository were already well-annotated, but that was not the case for the NRPS BGCs predicted by antiSMASH. In order to carefully identify duplicates among the group of predicted BGCs, we developed an expedited homology comparison based on cluster-prints. A cluster-print is a string representation of a BGC where each character (separated by a comma) refers to a specific

NRPS domain and hyphens are used as a delimiter to distinguish between different NRPS polypeptides. This method permits BGCs to be quickly compared to one another while avoiding complex sequence comparisons. For example, the cluster-print for tyrocidine, an NRP from *Bacillus brevis*, would be [A, T, E, -, C, A, T, C, A, T, C, A, T, E, -, C, A, T, C, A, T, C, A, T, C, A, T, C, A, T, C, A, T, C, A, T, -, T]. When two BGCs show identical cluster prints, we then compare the sequence of their predicted activated substrates. For example, for tyrocidine this would be [dPhe—Pro, Phe, dPhe, Asn—Gln, Tyr, Val, Orn, Leu]. We were thus able to determine how many unique NRPs a single IML was associated with (Supplementary File S4).

Our analysis has revealed the presence of 2703 IMLs that were involved in the biosynthesis of at least two or more distinct NRP products based on their cluster-prints. For instance, the IML 'Gly-LAPAAQGGIVRCARDA-Thr' was found in 90 distinct NRP products across three different species. When we conducted the same analysis using the well-annotated NRPs from the MIBiG repository, we similarly observed that some IMLs are involved in generating distinct NRP products. For instance, the BGCs of syringomycin and syringopeptin, both produced by *Pseudomonas syringae*, share multiple identical IMLs. Moreover, we also observed that highly similar linkers with more than 90% similarity are involved in the biosynthesis of distinct NRP products. For example, the IML 'Ile-AGRSSLPPIVPVSR-Nrp' is involved in the biosynthesis of sessilin (produced by *Pseudomonas sp*. CMR12a), while the IML 'Leu-AGRSSLPPILPVSR-Nrp' is involved in the biosynthesis of tolaasin (produced by *Pseudomonas costantinii*). These results not only indicate that highly similar to identical module-specific IMLs can be utilized to generate distinct NRPs, but also validates the application of our cluster-print approach to detect distinct NRP-generating BGCs and reflect the major role HGT play in bacterial evolution.

4 Discussion

There are two major resources hosting well-identified NRPs: (i) the NORINE database (Caboche *et al.*, 2008) and (ii) the MIBIG repository (Medema *et al.*, 2015). The former includes 1187 NRPs, while the latter contains 433 NRPs. However, when it comes to putative NRPS BGCs, our study comprises a total of 51 810 potential NRPS clusters (Supplementary Table S2), from which 7441 are identified as completely assembled NRPS clusters. We defined complete clusters as those possessing at least three modules and two IMLs. To the best of our knowledge this is the largest number of putative NRPS BGCs predicted from known genomic databases (39 232 bacterial genomes). Other studies have reported only 6351 (Dejong *et al.*, 2016) and 1704 (Cimermancic *et al.*, 2014) NRPS clusters.

All of the extracted NRPS BGCs were classified into 7365 unique cluster-prints, each of which potentially generates a novel NRP (Supplementary Fig. S10A). If so, the inclusion of the additional genomes results in a 27-fold increase in potential NRPs compared to those captured in the MIBiG repository. This increase likely reflects the fact that the MIBiG repository is based on 243 unique bacterial strains, while our more complete genome analysis comprised over 31 338 unique bacterial strains (based on their NCBI taxonomy identity designator) (Supplementary Fig. S10B). The large number of bacterial genomes processed in our study and the many NRPS BGCs identified using antiSMASH will certainly help the community by revealing potentially novel, not-yet-annotated NRPS BGCs. We are confident that a similar increase in NRPSs would result from expanding the scope of this analysis to include fungi and marine plants.

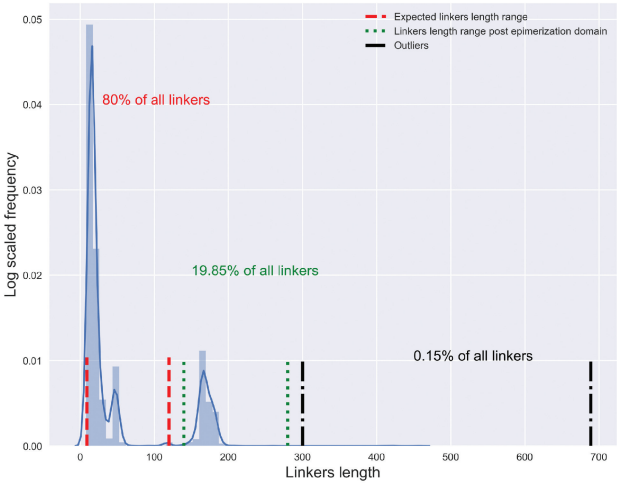

Fig. 4. Linkers length distribution: There are three clusters of IMLs based on their lengths: Linkers with lengths ranging between 9 and 120 amino acids. Linkers with lengths ranging between 160 and 280 amino acids. These are linkers with more than 300 amino acids in length (outliers)

IMLs could be clustered into three clusters based on their lengths (Fig. 4): (i) Linkers with lengths ranging between 9 and 120 amino acids. These are typical lengths and they represent 80% of all IMLs. (ii) Linkers with lengths ranging between 160 and 280 amino acids. These are linkers that succeed an epimerization domain in a BGC and they represent 19.85% of all linkers. These seem longer, due to only-recently-annotated domain 'TIGR01720' in TIGRFAMs protein family (Haft *et al.*, 2001), which is located immediately downstream of the epimerization domain and upstream of the condensation domain of the successive module. (iii) Linkers with length >300 amino acids. These are most certainly outliers and they represent less than 0.15% of all linkers. The genesis of these outliers is due to the limitation that antiSMASH tool sometimes has in properly defining border domains in case of new yet undefined and unannotated domains.

The identification of borders between distinct domains is crucial for our analysis. IML is the linker region between two successive modules. Precisely, the region between the peptidyl carrier protein domain (PCP or T-domain) of the first module and the condensation domain (C-domain) of the successive module. Thus, determining where the T-domain ends and the C-domain begins is vital for extracting the right linker region. Unfortunately, there is a lack of multi-modular crystal structures for NRPSs. Therefore, we used antiSMASH to predict all potential NRP BGCs, from which we extracted all our IMLs. antiSMASH depends on pHMMs to predict domains borders in a BGC. Hence, the quality of our extracted linker regions is only as good as the antiSMASH domain border identification algorithm. Fortunately, these pHMMs are constantly being updated and re-trained, allowing for improved predictive power as new data and new annotated domains are added.

Historically the importance of IMLs compatibility with adjacent modules has not been considered during NRP biosynthesis strategies. For instance, Nguyen *et al.* (2006) have applied several combinatorial biosynthesis strategies to produce a library of daptomycin analogues. Among other approaches the authors replaced entire modules within NRPS subunits. Here, we will focus on the module replacement strategy, and the role that the IML considerations might have played in their results. All the derived peptides are based on daptomycin, a cyclic 13-amino acid lipopeptide obtained from

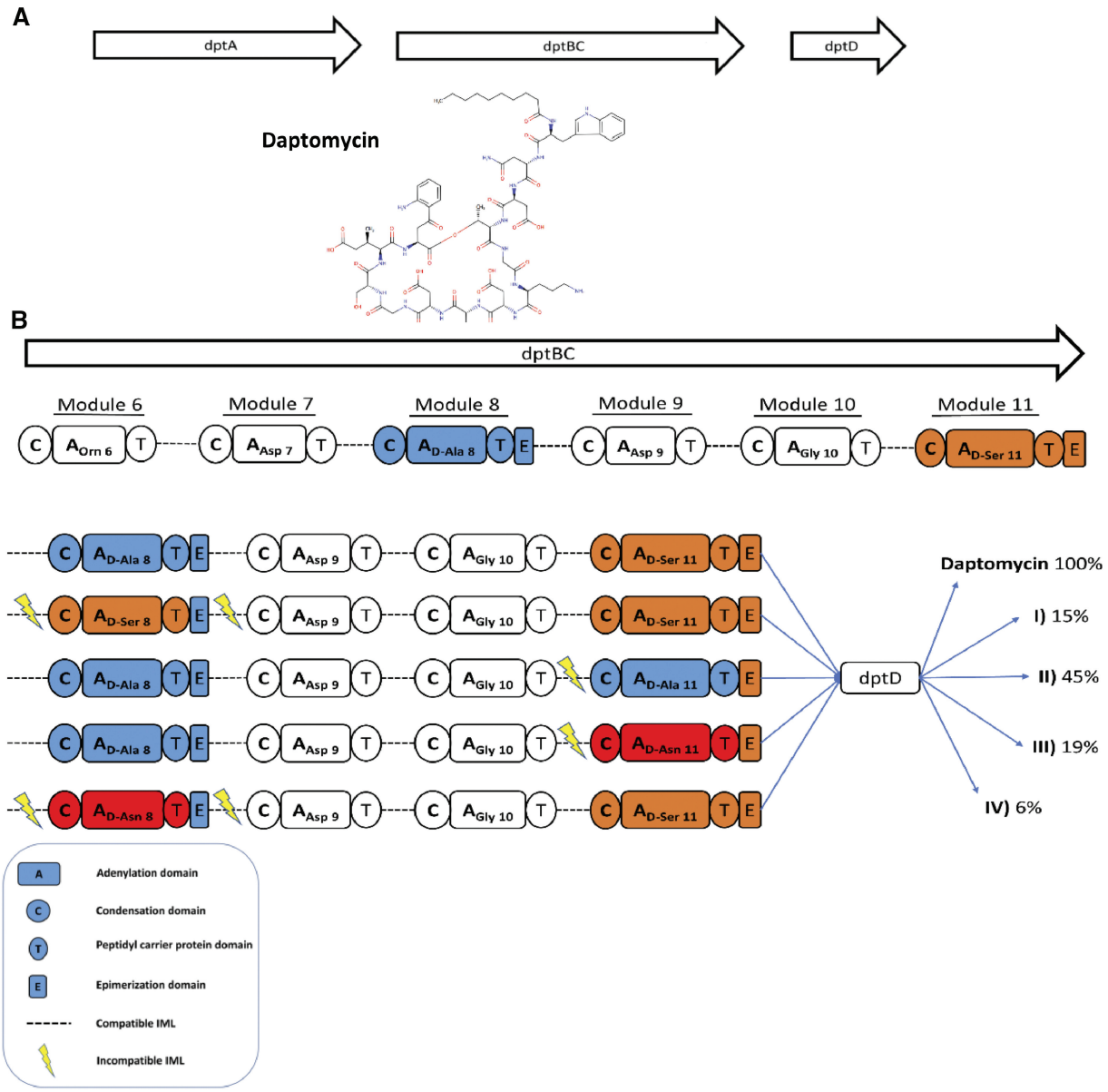

Fig. 5. Retrospective analysis of daptomycin analogues biosynthesis: (**A**) Daptomycin BGC from *Streptomyces roseosporus*. (**B**) NRPS organization of the daptomycin cluster and schematic showing module exchange strategy. Modules 8 and 11 were swapped for each other, or for the Asn 11 module from A54145 biosynthesis from *Streptomyces fradiae*. The swapping resulted in the synthesis of four daptomycin analogues I, II, III and IV each with 15, 45, 19 and 6% yield relative to the wild-type. The lightning bolts signify IML incompatibility

Streptomyces roseosporus that is a product of three biosynthetic NRPS subunits, dptA, dptBC and dptD (Fig. 5A).

Two strategies were conducted by Nguyen *et al.* (2006): (i) Exchange of homologous modules within the dptBC NRPS subunit. The other one was to undergo an (ii) Exchange of single heterologous modules.

Exchange of homologous modules within dptBC: Here, Nguyen *et al.* (2006) decided to conduct two experiments that involve replacing entire modules (C-A-T) within the dptBC NRPS subunit. Module 8 and module 11, which activate D-Ala8 and D-Ser11, respectively, were replaced. (I) The D-alanine encoding C-A-T from module 8 was deleted and replaced with the C-A-T from module 11 (change of Ala8 to Ser8). (II) The opposite replacement was also made where the C-A-T from module 11 was replaced with the C-A-T from module 8 (change of Ser11 to Ala11). The E domains of each module were left intact in an attempt to preserve the downstream inter-module associations. Production of the predicted D-Ser8 and D-Ala11 containing daptomycin analogues was observed, albeit at reduced production levels of approximately 15 and 45% relative to wild-type. The authors reasoned that the success of synthesizing those daptomycin analogues was due to the fact that both modules are highly homologous. However, the authors failed to explain why the yields were much lower than the wild-type and why the yield of (I) was lower relative to (II) (Fig. 5B).

We hypothesize that a possible reason for these decreased yields is due to IML incompatibility after module replacement. In the first experiment a middle module was replaced, giving rise to two incompatible IMLs (one on each side of the replaced module). In the second experiment, a terminal module was replaced, causing a single incompatible IML. Thus, the yield was 15% and 45% for the first and the second experiment, respectively.

Exchange of single heterologous modules: Here, module 11, which is selective towards D-Asn11, was obtained from the A54145 BGC from *Streptomyces fradiae*. The extracted module was used to replace either D-Ala8 or D-Ser11 from dptBC NRPS subunit of the daptomycin BGC. This approach led to the isolation of two new analogues [D-Asn11 (III) and D-Asn8 (IV)]; however, yields were even further reduced relative to wild type, i.e. 19 and 6%, respectively (Fig. 5B).

We again hypothesize that the main reason for such a steep drop in the yield is due to the impact of incompatible IMLs post module replacement. Moreover, we know from our analysis that IMLs are not conserved across species and thus replacing modules across species would further increase the level of IML incompatibility and would result in a more pronounced effect on the product yield. Similar analysis was conducted on the findings of Bozhüyük *et al.* (2018) (Supplementary Information Retrospective analysis). These observations show that the compatibility of the IML with the entire adjacent A-domain is critical to ensure a proper yield of the NRP product. These data support the idea that module-specific IMLs are critical to the successful generation of NRPs.

5 Conclusion

Using our IML NRPS-Parser, we extracted more than 39k NRPS IMLs and analyzed their association with their adjacent A domain substrates. This led to the discovery that IMLs are very specific to the A domain modules that they connect, with more than 92% of the identified IMLs being associated with a specific pair of modules. We also determined that the same IML could be involved in the biosynthesis of different NRP products across various bacterial genera (Supplementary File S4). Overall, however, IMLs that link a particular module pair show a low degree of conservation across bacterial genera. We also determined that IMLs exhibit more secondary structures (α-helices) than IDLs, however, they share similar hydrophobic profile. Furthermore, as a proof-of-concept, we retrospectively analyzed the findings of (Nguyen *et al.*, 2006) and (Bozhüyük *et al.*, 2018) demonstrating that IMLs incompatibility could dramatically impact biosynthetic yields of daptomycin lipopeptides and ambactin analogues. Overall, our data indicate a strong relationship between NRPS IMLs and their adjacent A domains. This finding suggests that, going forward, combinatorial biosynthesis strategies to generate novel NRPs should consider IMLs in addition to other established parameters (Baltz *et al.*, 2006; Bozhüyük *et al.*, 2018; Calcott *et al.*, 2014; Coëffet-Le Gal *et al.*, 2006; Crüsemann *et al.*, 2013; Meyer *et al.*, 2016; Nguyen *et al.*, 2006).

All 39 804 IMLs extracted in this study (Supplementary Table S2) as well as our parser are publicly available at https://nrps-linker.unc.edu/. We anticipate this tool will not only facilitate mining the data we have analyzed here, but will also enable interested researchers to expand their studies as new genomes (bacterial, fungal and plant) are obtained. Our study lays the foundation for future experimental validations of our hypothesis that IMLs play a crucial role in governing the biosynthesis of NRPs. We expect that additional approaches and tools could be developed that rely on this finding and facilitate the design of novel NRPS BGCs using the most appropriate IMLs for combinatorial biosynthesis of novel NRPs.

Acknowledgments

We thank Vladimir Jojic, Iva Farag, Ammar Abdulmughni and Stephen J. Capuzzi for their useful comments and their constructive suggestions. We would also like to thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

Funding

This work was supported by the Eshelman Institute for Innovation grant [R1010 RX03620204 to Sherif Farag], and the National Institutes of Health (GM112981 to E. A. S.).

Conflict of Interest: none declared.

References

Bae,K. *et al.* (2005) Prediction of protein interdomain linker regions by a hidden Markov model. *Bioinformatics*, **21**, 2264–2270.

Baltz,R.H. (2006) Molecular engineering approaches to peptide, polyketide and other antibiotics. *Nat. Biotechnol.*, **24**, 1533–1540.

Baltz,R.H. *et al.* (2006) Combinatorial biosynthesis of lipopeptide antibiotics in Streptomyces roseosporus. *J. Ind. Microbiol. Biotechnol.*, **33**, 66–74.

Beer,R. *et al.* (2014) Creating functional engineered variants of the single-module non-ribosomal peptide synthetase IndC by T domain exchange. *Mol. BioSyst.*, **10**, 1709–1718.

Bhaskara,R.M. *et al.* (2013) Understanding the role of domain-domain linkers in the spatial orientation of domains in multi-domain proteins. *J. Biomol. Struct. Dyn.*, **31**, 1467–1480.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.

Bozhüyük,K.A.J. *et al.* (2018) De novo design and engineering of non-ribosomal peptide synthetases. *Nat. Chem.*, **10**, 275–281.

Caboche,S. *et al.* (2008) NORINE: a database of nonribosomal peptides, Nucleic acids research, **36** (Database issue), D326–D331.

Calcott,M.J. *et al.* (2014) Biosynthesis of novel pyoverdines by domain substitution in a nonribosomal peptide synthetase of pseudomonas aeruginosa. *Appl. Environ. Microbiol.*, **80**, 5723–5731.

Cimermancic,P. *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, **158**, 412–421.

Coëffet-Le Gal,M.F. *et al.* (2006) Complementation of daptomycin dptA and dptD deletion mutations in trans and production of hybrid lipopeptide antibiotics. *Microbiology*, **152**, 2993–3001.

Crüsemann,M. *et al.* (2013) Evolution-guided engineering of nonribosomal peptide synthetase adenylation domains. *Chem. Sci.*, **4**, 1041.

Dejong,C.A. *et al.* (2016) Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.*, **12**, 1007–1014.

Doekel,S. *et al.* (2008) Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in Streptomyces roseosporus. *Microbiology*, **154**, 2872–2880.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Felnagle,E.A. *et al.* (2008) Production of medically relevant natural products nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Mol. Pharm.*, **5**, 191–211.

George,R.A. and Heringa,J. (2002) An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng. Des. Sel.*, **15**, 871–879.

Gokhale,R.S. and Khosla,C. (2000) Role of linkers in communication between protein modules. *Curr. Opin. Chem. Biol.*, **4**, 22–27.

Haft,D. *et al.* (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **29**, 41–43.

Lott,J.S. and Lee,T.V. (2017) Revealing the inter-module interactions of multi-modular nonribosomal peptide synthetases. *Structure*, **25**, 693–695.

Medema,M.H. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.

Meyer,S. *et al.* (2016) Biochemical dissection of the natural diversification of microcystin provides lessons for synthetic biology of NRPS. *Cell Chem. Biol.*, **23**, 462–471.

Mootz,H.D. *et al.* (2000) Construction of hybrid peptide synthetases by module and domain fusions. *Proc. Natl. Acad. Sci. USA*, **97**, 5848–5853.

Nguyen,K.T. *et al.* (2006) Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proc. Natl. Acad. Sci. USA*, **103**, 17462–17467.

Reger,A.S. *et al*. (2007) Biochemical and crystallographic analysis of substrate binding and conformational changes in acetyl-CoA synthetase. *Biochemistry*, **46**, 6536–6546.

Robinson,C.R. and Sauer,R.T. (1998) Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. *Proc. Natl. Acad. Sci. USA*, **95**, 5929–5934.

Stevens,B.W. *et al*. (2005) Progress toward re-engineering non-ribosomal peptide synthetase proteins: a potential new source of pharmacological agents. *Drug Dev. Res*., **66**, 9–18.

Tarry,M.J. *et al*. (2017) X-ray crystallography and electron microscopy of cross- and multi-module nonribosomal peptide synthetase proteins reveal a flexible architecture. *Structure*, **25**, 783–793.e4.

Udwary,D.W. *et al*. (2002) A method for prediction of the locations of linker regions within large multifunctional proteins, and application to a type I polyketide synthase. *J. Mol. Biol*., **323**, 585–598.

Weber,T. *et al*. (2015) antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res*., **43**, W237–W243.

Winn,M. *et al*. (2016) Recent advances in engineering nonribosomal peptide assembly lines. *Nat. Prod. Rep*., **33**, 317–347.

Wu,R. *et al*. (2009) The mechanism of domain alternation in the acyl-adenylate forming ligase superfamily member 4-chlorobenzoate: coenzyme A ligase. *Biochemistry*, **48**, 4115–4125.

Yu,D. *et al*. (2013) Functional dissection and module swapping of fungal cyclooligomer depsipeptide synthetases. *Chem. Commun*., **49**, 6176.