Non-ribosomal peptide synthetase domain boundary identification and new motifs discovery based on motif-intermotifs standardized architecture

Ruolin He,[1] Jinyu Zhang,[2,3] Yuanzhe Shao,[4] Shaohua Gu,[1,4] Wen-Bing Yin,[2,3,*] and Zhiyuan Li[1,4,*]

[1] Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[2] State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, PR China

[3] Savaid Medical School, University of Chinese Academy of Sciences, Beijing, 100049, PR China

[4] Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

* To whom correspondence should be addressed. Tel: +86 10 62750190; Email: zhiyuanli@pku.edu.cn

Correspondence may also be addressed to Wen-Bing Yin. Tel: +86 10 64806170; Email: yinwb@im.ac.cn

Abstract

Non-ribosomal peptide synthetase (NRPS) is a diverse family of biosynthetic enzymes for the assembly of bioactive peptides. Its repetitive modular structure inspired many reengineering efforts. Despite the continued expansion of microbial sequencing data, varied standards for annotating domains and modules make data-driven discoveries in NRPSs challenging. In this work, we presented a motif-and-intermotif architecture of NRPSs by partitioning the condensation (C), adenylation (A), thiolation (T) and thioesterase (TE) domains by conserved core motifs. This "standardization" enables systematic evaluation of functional couplings from a large number of NRPS sequences, resulting in the discovery of novel motifs such as glycine (G)-motif with functional implications in structural flexibility. The importance of this motif was verified by mutation experiments in an antitumor fumiquinazoline pathway. Additionally, coevolution studies clarify significant barriers associated with reengineering NRPSs, and comparison of sequences between species reveals entanglement between phylogeny and substrate specificity. Our findings give an in-depth analysis and statistical insights into the sequencing characteristics of NRPSs, paving the path for future data-driven discoveries.

Introduction

Non-ribosomal peptide (NRP) synthesized by non-ribosomal peptide synthetase (NRPS) is a diverse family of natural products widespread in fungi and bacteria(1-3). According to a recent survey, NRPS is the most prevalent class (25.45%) out of the 1,096,473 biosynthetic gene clusters (BGCs) detected from 18,661 bacterial genomes, and 42.58% out of the 124,092 BGCs from fungal genomes(4). Microbes exploit NRP for various niche-construction activities, such as communication (5,6), defense(7-11) and resource-scavenging(12). Diverse NRP serve as reservoirs for pharmaceutical innovations, with anti-cancer, anti-fungal, anti-viral, cytotoxic, and immunosuppressive activities(13-15). To date, nearly thirty NRP structural medicines have been approved for commercialization(13), including actinomycin(16), penicillin(17,18) and vancomycin(19).

Many reengineering efforts have been inspired by the pharmacological potentials of NRPS, as well as its modular architecture (20,21). NRPS synthesizes peptides in an assembly-line fashion using its repeating module units. Each module is generally comprised of three key domains, including the adenylation (A) domain, which recognizes the substrate and activates it as its aminoacyl adenylate; the condensation (C) domain, which catalyzes amide bond formation between two substrates; and the thiolation (T) domain, which shuttles the substrates and peptide intermediates between catalytic domains. Other optional domains are also important, such as the thioesterase (TE) domain or terminal condensation-like domain (C_T)domain(22), which cyclize or release peptide from the NRPS and generally occurs at the end of a NRPS; epimerization (E) domain, which catalyzes conversion of L-amino acids substrate to D-amino acids and usually follows the C domain in case it occurs. In order to create novel products, many synthetic efforts have been made to re-arrange these building blocks of NRPSs. For example, due to the central roles of A domains in substrate recognition and activation, early works focused on the A domain, including A domain substitution, reprogramming A domain substrate coding residues, and A subdomain substitution(20). Later, considering possible roles of C domains in substrate selectivity, researchers attempted to re-engineer catalytic domains together with the A domain, such as substituting a complete module, or the C-A domain and T-C-A domain(20). Recently, Bozhüyük et al. forward the exchange unit concept, and recombined NRPS at a fusion point between C and A domain where they named as “C-A linker”(23) , and they also found another cutting point inside the C domain(24). Calcott et al. also attempted to recombine at a another cutting point between C and A domain(25), and this point was almost located the start of C-A linker defined by Bozhüyük et al. However, in most cases, reengineering would substantially reduce the product yield. Furthermore, these unit’s conversions were performed between two substrates with similar chemical properties, based on well-studied NRPS systems with a small number of sequences(26,27). A systematic understanding towards the rational design of NRPS has yet to be developed.

The massive growth of the microbial sequencing data has created opportunities towards a comprehensive understanding of NRPSs. As of October 2021, the Integrated Microbial Genomes Atlas of Biosynthetic gene Clusters (IMG-ABC) v5 database contains a total of

411,027 BGCs based on sequence annotation(28), and Minimum Information about a Biosynthetic Gene Cluster (MiBiG) v2 database contains 1926 BGCs confirmed to be biologically active(29). Together with accelerated expansion of databases, bioinformatics tools and platforms that specifically predict and annotate the secondary metabolites are also in rapid development(30), such as antiSMASH(31), PRISM(32), SeMPI(33). Pfam, a widely utilized database of protein families and domains since 1997(34,35), have also been commonly used for NRPS domain detection by profile hidden markov models (HMM). However, different bioinformatics tools and platforms employ different annotation standards and algorithms, and their definitions of domains and inter-domains in NRPS are perplexing.

Aside from annotating the module-domain architecture, many algorithms have been dedicated to predict the substrate specificity of the A domain based on its sequence(33,36-39). Some of these methods are entirely supervised learning-based, with full-length A domain sequences and their proven substrates serving as the training set (37). Some of the longest-developped methods also incooperated the structural information of the A domain (40). With advances in protein structures, the substrate-specifying residues of modular mega-enzymes, such as NRPS(26,41) and polyketide synthase (PKS)(22), were identified. Based on the first crystal structure of A domain (PheA, PDB 1amu)(26), Stachelhaus et al. identified ten substrate-specifying residues in the A domain's binding pocket. These ten residues strongly associate with the substrate specificity, and have been widely exploited by researchers to improve prediction efficiency (40). For simplicity, in the following text we refer to these ten substrate-specifying residues as the 10-S code ('S' is the abbreviation for Stachelhaus). Despite the fact that the 10-S code is highly predictive of substrate specificity, the substrate seems unable to determine the 10-S code. Actually, diverse forms of the 10-S code could exist for the same substrate(40). Such diversity of 10-S code causes barriers in extending the prediction algorithm: without a one-to-one mapping between 10-S code and substrate specificity, the performance of current machine learning algorithms heavily replies on the training set. Actually, performances of these algorithms in fungal NRPSs are unsatisfactory, as fungi have a significantly smaller number of A domains with known substrate. Furthermore, reengineering based on modifying the 10-S code does not operate as expected(41), highlighting the complexities in the A domain's function.

Coevolution analysis has been used as a general strategy to uncovering protein interactions and functional couplings from a large number of sequences (42). Existing methods in coevolution analysis include normalized mutual information (MI) of amino acid frequencies(43), direct coupling analysis (DCA)(44), protein sparse inverse covariance (PSICOV)(45) and statistical coupling analysis (SCA)(46). Among them, SCA has been utilized to identify appropriate boundaries for protein reengineering(47). The basic idea of SCA is to identify groups of residues sharing the same modes of covariations, which are referred to as "sectors" (46). Residues projecting to a single sector can be viewed as "evolving together", and are likely to cooperate in function (48). Therefore, sectors could serve as the "evolutionary units" for reengineering, as borders of sectors provide rational

cutting-points for preserving functional connection between residues (46).

Applying SCA to NRPS sequences to gain insights on reengineering is a straightforward idea. However, one obstacle in this simple notion is the paucity of manually curated sequences: the majority of sequences in databases were automatically annotated by machine learning algorithms, and have not yet been reviewed by human experts. Moreover, different algorithms with varying standards for annotation were adopted by different researchers, which muddled the exact definitions of domains and interdomains in NRPSs. For examples, A and C domains annotated by the default option of antiSMASH are substantially shorter than Pfam-annotated domains (34) , and are usually different with sequences from the Protein Data Bank (PDB) (49-51). Even the domain annotations from antiSMASH, one of the most popular annotation platforms for NPRSs, change slightly between versions.

One solution for this "confusion by the annotation" is to focus on the invariant sequence pieces in NRPSs architecture, known as “core motifs” (52). Despite the enormous diversity in NRPs, highly conserved sequence elements have been discovered across domains, usually essential to the domain function. These well-established core motifs include: 7 motifs in the C domain, 10 motifs in the A domain (of note, these are not the substrate-specifying 10-S code), 1 motif in the T domain and 1 motif in TE domain(52). These conserved motifs could be used to standardize NRPS domains, allowing for statistical analysis of NRPS sequences from large databases without manual curation.

In our work, we constructed a “coordinate system” for NRPSs based on these previously well-established core motifs. Such motif-and-intermotif architecture extracted global statistics on the length, variation, and function-related information in every piece of the NRPS sequence from several databases. Following that, we were able to systematically evaluate function couplings from a large number of sequences. For example, we uncovered several previously unrecognized conserved motifs, which could have functional implications as well as serve as candidates for novel core motifs. One of these novel motifs is located close to the A domains specificity pocket. O ur computational analysis predicted it being crucial for NRPS structural flexibility during peptide synth esis, which was then supported by mutation experiments. Then, using coevolution analysis, we dissected two major challenges in reengineering NRPSs: first, there is no cutting point to separate multiple overlapping coevolving sectors, limiting domain/subdomain swapping; second, the specificity of the A domain intertwines with not only the classical 10-S code but also the length of five loop regions in the A3-G intermotif, complicating the rational modification of A domain. We also assessed the differences between bacteria and fungi by this standardized architecture, and identified several inter-motifs that significantly diverge in length, further emphasizing the importance of NRPS architecture. Our findings provide a comprehensive overview and statistical insights into the sequence features of NRPSs, paving the way for future data-driven discoveries. In the end, an online platform entitled "NRPS Motif Finder" is offered for parsing the motif-and-intermotif architecture (http://www.bdainformatics.org/page?type=NRPSMotifFinder).

Result

1. Overview of the motif-and-intermotif architecture of NRPS

We first partitioned the NRPSs architecture based on well-known core motifs. The sequence representations of 19 well-studied motifs were curated from literature(52), then located to 7329 NRPs domains from MiBiG database by alignment (see Methods for detail). As expected, most of the known core motifs are highly conserved, exhibiting clear sequence logos (first panel, Fig1 A-D, Fig S2-6). Motifs from A, E, and T domains have sequence identities higher than >60%. The sequence identities of motifs from C and TE domains are slightly lower, but also around 40%, except motifs C6 and C7 (identities 22% and 28%, respectively). On contrary, sequence regions between motifs (termed as “intermotif region”) have substantially lower sequence identities, ranging from 10%-25%. The length distributions of most of the intermotif regions are generally narrow with a standard deviation smaller than 10 aa (third panel, Fig1 A-C), demonstrating the architecture invariance of NRPSs. Yet, unlike other domains, the length from the TE motif to the end of the cluster has a much broader distribution (320-380aa, Fig1 D). Of note, the antiSMASH-annotated C domain by its default option doesnt include the motif C6 and C7; antiSMASH-annotated A domain doesn’t include the motif A9 and A10 by default; and the antiSMASH-annotated E domain doesn’t include the motif E6 and E7 by default (second panel, Fig1 A-C).

Overall, such conserved-motif and variable-intermotif architecture support the use of known core motifs as "coordinates" of NPRSs. This standardization enables statistical characterization on the sequence-function connections. In addition, we constructed an online platform “NRPS Motif Finder” (http://www.bdainformatics.org/page?type=NRPSMotifFinder) for parsing the motif-and-intermotif standardized architecture of NRPS from its coding sequences.

1.1. Sequence properties related to C subtypes

As descripted in Rausch. et al(53), C domains contain various functional subtypes (such as Dual E/C,LCL, DCL, Starter, and Heterocyclization), which are mostly related to the chirality of substrates. Different subtypes have distinctive sequences(53), which may explain the relatively low sequence identity of C domain motifs (Fig 1A). By extracting the known motifs from 1765 C domain sequences from the MiBiG database, we analyzed the motif logos for different C domain subtypes (Fig S2). The motif C2 and C3 are the most invariant across all C subtypes (Fig 1A, first panel), with identity 49% and 55% respectively. Motif C6 and C7 are the least conserved even within each subtype (Fig 1A and Fig S2). Motif C3 was reported to be crucial for catalysis(53,54). It is worth noting that the motif C3 in the Heterocyclization subtype is DxxxxD, as distinguishable from the conserved HHxxxD in other subtypes. In general, motifs from the DCL and the Starter subtype are more variable, while motifs from the Dual E/C subtype are conserved and distinctive from other subtypes, making it possible to recognize Dual E/C subtype merely based on motifs.

Although the motif of T (T1) is highly conserved in sequence (75% identity), the intermotif region between the motif T1 and the next motif C1 exhibit a bi-modal distribution (Fig 1B). A detailed examination of the domain annotation reveals that the length of T1-C1 region depends on the annotation of the following C domain (Fig S4B): regions followed by the LCL subtype are of length 65-70 aa, whereas regions followed by the Dual subtype are of length 95-105 aa. This observation suggests that the architecture of intermotif regions may have functional implications.

1.2. Mutual information between sequence and functional subtypes in C domain and A domain.

Since C domain sequences have been demonstrated to correlate with chirality subtypes, we investigated the Shannon information between sequences and C domain functional subtypes (see method). The region between motifs C1 and C7, as represented in the last panel of Fig 1A, contains a high content of information about the chirality subtypes. Following the C7 motif, the chirality information gradually decreases until it reaches the background level as the sequence approaches the A1 motif. (Fig 1A).

Given the critical role of the A domain in substrate selectivity, we employed the same mutual information measurement to ascertain the degree to which each A domain position is related with substrate specificity (see method). As seen in the last panel of Fig 1B, high-information positions begin with motif A3 and finish before motif A6. This region overlaps the substrates binding pocket (49). The A4-A5 region is well-known for containing the substrate-predicting 10-S code (26).

2. Discovery and functional identification of new conserved motifs

We calculated the amino acid frequency and gap frequency of the module alignment using the aforementioned standards on the sequences of 1161 (1053 from bacteria, 81 from fungi and 27 from other) C-A-T NRPS modules from the MiBiG database (Fig 2A, lower panel). As expected, the majority of highly conserved positions were found within or right beside known core motifs. These highly conserved positions could be seen as the extension of known motifs (Fig S7). However, there are four highly conserved positions relatively distant from well-known motifs (Fig 2A, upper panel), and hence cannot be regarded as extensions of these well-known motifs.

2.1. The new “G-motif” in the substrate binding pocket mediates conformation flexibility

A conserved glycine located in the middle of the A5-A6 intermotif region attracts our attention. It is 21(±2) aa from the last residue of the A5 motif, and 23(±0.2) aa from the first residue of the A6 motif. This glycine is found to be highly conserved in more than 99% of the aligned A domains, with surrounding residues being moderately conserved (the second logo in the upper most panel of Fig 2A). Notably, this glycine divides the A5-A6 intermotif region into two halves in terms of substrates information, where the sequences preceding

this Gly contains substantially more mutual information about substrate specificity (Fig S8). For convenience, well refer to this conserved glycine and its surrounding residues together as the "G-motif".

We then located this G-motif to the G505 position in a recently solved NRPS structure, the linear gramicidin synthetase subunit A (LgrA) (55). To our surprise, the G-motif locates on a loop structure, which is usually considered non-conserved. The LgrA structure has been solved with three function-related states: 6MFY, the substrate donation state; 6MG0, the thiolation state; 6MFZ, the condensation state (55). By calculating hydrogen bonds on these structures, we noticed that the G-motif (I502, P503, I504, G505, K506, P507 and I508) interact with N479 in the A5 motif (where A5 located to: N479-E485, sequence: NGYGPTE) and its back side residues (V490, T491 and M492). It also interacts with the back side residues of the A7 motif (D575 and I578; with A7 located to Y564-R569, sequence: YRTGDR) (Fig 2B-D). Of note, number of hydrogen bonds for these interactions change remarkably in different function-related states (Fig 2B-D, Fig S9). In the substrate donation state 6MFY, the G-motif has 12 hydrogen bonds (average bond length 3.13±0.27Å); The thiolation state 6MG0 has 9 bonds (average length 3.09±0.24Å); The condensation state 6MFZ has only 6 bonds (average length 3.13±0.27Å).

We hypothesized that the interactions indicated above are functionally significant. One prediction is that, in addition to the highly conserved glycine, the chemical properties of other residues involved in these interactions should be conserved. Therefore, we turned back to the 1161 MiBiG C-A-T alignment to checked the chemical conservation of the residues interacting with the G-motif (Table S1). The cross-species conservation of chemical properties agrees with our prediction (Table 1). Of note, in these residues, only T491 and N479 use the hydroxyl or amide group in the side chain to form hydrogen bonds. Other hydrogen bonds, on the other hand, are formed by the common α-carboxyl group and α-amino group in the main chains. The fact that most of these residues are still conserved suggests causes for the conservation is beyond forming hydrogen.

By simulated mutagenesis, we found that the position of glycine in the G-motif only permits small amino acids such as glycine and alanine. Mutating this residue into amino acids with larger volume or with inflexible loops resulted in a disturbance of surrounding structures (Fig 3A-B).

In order to verify the function of this highly conserved glycine, we performed mutation experiment of “G-motif” in a monomodular NRPS FmqC involving in the biosynthesis of fumiquinazoline C (FQC) in *Aspergillus fumigatus* (Fig 3C). In this system, a tripeptide precursor fumiquinazoline F (FQF) was first synthesized by FmqA and then the indole side chain of FQF oxidized by the FmqB. Then, the monomodular NRPS FmqC, activated L-Ala to create the product of fumiquinazoline A (FQA), thereby constructing new C_2–N bond and finally forms FQC(56) (Fig 3D). The *fmqC* gene deletion resulted in the disappearance of FQC as the pathway diverged towards a putative compound (termed as compound **1**), as indicated in the first two rows of Fig S10. This is consistent with previous study in *A.*

fumtigatus AF293(57). We reintroduced *fmqC* into the original locus in the △*fmqC* background and generated one control strain with full-length copy of *fmqC* in *A. fumtigatus*. The control strain could produce FQC just as the wild type (Fig S10). Then, the glycine in the G-motif of FmqC (G409) was mutated to six other representative amino acids (A, alanine, R, arginine, D, aspartic acid, P, proline, W, tryptophan and Y, tyrosine), to generate six FmqC (G409) mutants in *A. fumtigatus*, respectively. Following LC-MS analysis, the FQC yield of all mutants is lower than control. While the G-motif mutant G409A or G409R didn't strongly reduce the yield of FQC, the other four mutants substantially reduce the yield of FQC (Fig 3E-F). For example, when the glycine was mutated into proline (G409P), the yield of FQC was the only ~3% of the control (Fig 3F and Table S2). Although we observed no accumulation of the precursor FQF in strains with substantially decreased FQC, we detected compound **1** as the divergent product in all of these strains instead (Fig 3D-F). Our mutation experiment supported the critical role of the conserved glycine in G-motif, and indicated the flexibility of glycine is important in this position.

2.2. Conserved motif at the start of A domain

Another conserved residue discovered via standardized multiple sequence alignment is an Alanine locating in the start of Pfam annotated A domain (Fig 2A, the first sequence logo at the first panel). It is 11(±0.4) aa before the A1 motif starts, and 124(±4) aa after the C7 motif ends. This residue is found at A221 in the LgrA sequence. By calculating hydrogen bonds based on the LgrA structures, we noticed that this conserved A221, V222 and another relatively conserved Proline (P217 in LgrA) interact with residues in the A1 motif (L229, T230, Y231 and K232, where A1 located to: 229-234 in LgrA, sequence: LTYKQL), as well as back side of residues from the A4 motif (L412 and I414; where A4 located to 395-398 in LgrA, sequence: FDGS). And the latter also have direct interaction (between A410, L406 and Y231. Fig S11-12). The chemical characteristics of residues anticipated to interact with the conserved alanine are conserved in the alignment of 1161 sequences, similar to the G-motif (Table S1). We termed this motif as Aα1, which marks the start of A domain.

2.3. Conserved motif at the first helix of T domain

The other two conserved resides outside of known motifs are Glutamic acid and Leucine, both of which are found at the start of the Pfam-annotated T domain (Fig 2A, the third sequence logo at the first panel). They were found at E700 and L711 in the LgrA sequence. The conserved Glutamic acid is 10(±0.04) aa from the conserved Leucine, which is 8(±0.6) aa before the first residue of the T motif. The T domain is known to be a distorted four-helix bundle with an extended loop between the first two helices(58). According to the LgrA structure, the E700 and L711 are located at the N and C terminus of the first helix of the T domain, respectively. Distinct from the G-motif and the conserved A221, we found no hydrogen bonds outside the interior of this alpha helix. Instead, the E700 and L711 and other conserved aliphatic and aromatic amino acids take part in formation of the hydrophobic core (Fig S13), agreeable with previous reports that the four helices were bound together by the hydrophobic interaction(59). We termed this motif as Tα1, which

marks the start of T domain.

3. SCA reveals overlapped co-evolving sectors across C-A-T module

After examining the sequence properties of the standardized C-A-T sequences position by position, we employed coevolution method to investigate the coupling between positions. Statistical coupling analysis (SCA), a commonly used technique in detecting “evolutionary units”, were applied the 1161 aligned-and-standardized C-A-T sequences from MiBiG. In the SCA result, only a few top modes (representing the collective covariations of a set of residues known as "sectors") have eigenvalues (measures of covariation magnitude) that are significantly larger than those obtained by random shuffling (26 out of 2560 total, Fig S14). The dominant first mode of SCA displays global correlations, which are commonly regarded as phylogenetic association and are always ignored in subsequent analysis (46). 122 positions in MSA show significant patterns of correlation, and can be separated into two distinct groups (Fig 4A, two groups labeled green and magenta). The green sector contains residues primarily in the C domain, the magenta sector contains residues spanning across the whole NRPS module, but none of its residues locate to the A domain pocket region (Fig 4A).

Then we searched for sectors that contained a high proportion of A domain residues. The sector containing positions densely covering the A domain pocket region appears till the sixth biggest mode (Fig 4A, red dots), which includes three residues in the 10-S code. While the green sector varies nearly independently from the magenta and the red sector, residues contributing to the red and the magenta sectors share a significant degree of covariation (Fig 4C), indicating the entanglement between substrate-specifying residues in the A domain and other regions of NRPSs.

In proteins reengineering, a sector is usually considered as a unit as its residues share the same evolutionary mode. As shown in the second panel of Fig 4A by black marks, Bozhüyük et al. cut the NRPS at the beginning of the red sector in their 2019 work of XUC (black triangle); Both Calcott et al. and Bozhüyük et al. in their 2018 work of XU cut NRPS after the ending of green sector (black dot and diamond, respectively).

In addition, we also analyzed 685 C-A-T-C sequences and 245 C-A-T-E sequences by SCA (Fig S15A-B). In both sets of sequences, we found consistent sectors (Fig S15). In C-A-T-C, the first two sectors locate to two C domains respectively, and the third sector spans across the four domains. In C-A-T-E, one sector locates to C domain, and another sector spans across the four domains. Intriguingly, the third sector spans across T and E domains, suggesting a possible cut point locating at the beginning of this sector, between the A and the T domains (Fig S15).

4. Factors influencing the mapping from the A domain sequence to its substrate specificity

4.1 Entanglement between substrate specificity and phylogenetic history

We proceeded the search for specificity-related positions using a modified SCA method, employing 2363 (2370 from bacteria, 215 from fungi and 51 from other) aligned and standardized A domain sequences connected to their experimentally confirmed substrates (see method). (Fig S16D). We primarily focused on modes that co-vary with the substrate column (Fig 5A). Residues from these substrate-related sectors are found throughout the A domain, with an enrichment in the A3-A6 pocket region and a sparse distribution in the A2-A3 region. The phylogenetic sector (Fig 5A, sector 1), in particular, receives the second-largest contribution from the substrate column. It has long been recognized that phylogenetic relatedness results in consistent correlations across the whole sequence. As a result, substrate specificity, which is highly linked to the 10-S code, is also linked to phylogenetic covariation throughout the A domain. This observation provides an explanation for the diversity of the 10-S code even for the same substrate.

4.2 Architecture of the substrate binding pocket relates to the diversity of 10-S code

Generally, SCA method gives higher weight to conserved positions(46). Interestingly, some substrate-related sectors tend to locate to highly variable regions (Fig 5A and Fig S17). In aligning to PDB structures 1amu and 4zxi, we observed that these variable regions represent five protein loops in the A domain binding pocket (Fig 5A, loops region marked by arrows between motif A3-A6): A3-A4, A4-S4, S4-S6, S6-A5 and A5-G. A_n is the *n*-th core motif of A domain. S_n is the *n*-th 10-S code. G is the highly conserved glycine in the G-motif aforementioned).

Constructing an efficient MSA from highly varied regions is challenging. To further understand the functions of these specificity-related loops, we compared their basic architectures, which are quantified by their length in the unit of amino acids. We estimated the Euclidean distances between pairs of A domains using the lengths of the five loops as a five-element vector. We observed that this loop-based distance is connected with, but not totally determined by, the similarity of sequence alignments and the phylogeny of A domains. (Fig 5B).

This loop-based distance matrix was clustered hierarchically, generating five groups based on their loop-length vectors (Fig S18). Except for substrates Ala, Dhb, and Aad, one substrate typically corresponds to multiple loop-length groups. Importantly, the diversity of the 10-S code is significantly reduced within each loop-length group. For example, for the substrate cysteine, the 10-S code in the first loop-length group is readily distinguishable from that in the third loop-length group, despite the fact that both groups are constituted of A domains from distant phylogenies. Statistically, the diversity of the 10-S code is reduced by identifying its loop-length vector group (Fig S19).

5. Comparison of NRPS architecture between bacteria and fungi

The relationship between loop length and substrate specificity highlights the importance of basic architecture in the sequence function. It has been reported that bacteria and fungi have different A domain pockets, which could explain why predicting substrate specificity in fungi is challenging (41). With the sequence standardization, the difference now can be discerned with more clarity.

The intermotif length of bacteria and fungus differ across the A domain, except for the A3-A4 intermotif (Fig 6A). The median length of fungal A domains is longer than bacteria in the intermotif A2-A3 by 4 aa, shorter in A4-A5 by 5 aa, and longer in A6-A7 by 7 aa. The largest difference occurs in the sub-domain of A, where the median of fungal intermotif A8-A9 is 20 aa longer than that of bacteria, and the median of the A10-T1 intermotif is 11 aa longer in fungus. All the differences described above is statistically significant. In a detailed check, a considerable portion of the differences are contributed by the fungal A domains with Aad-specificity, one of three precursors of beta-lactam (60,61). Yet, even excluding the Aad-specificity fungal A domains, the above differences are still substantial and does not change much by the median (Fig S20A-B).

While the fungal A domain tends to be longer as a whole, it has shorter loops within the A pocket region, regardless of whether the substrate is Aad (Fig 6B). Four of the five variable loops (A4-S4, S4-S6, S6-A5 and A5-G) have median lengths that are 1, 3, 1 and 1 aa shorter in fungus, respectively. The A3-A4 loop does not show statistically significant difference in length. Cumulatively, the median length of five loops in fungus is 6 aa shorter than bacteria by the median, supporting the hypothesis of Lee et al that the substrate-interacting locations of the fungal A domain differ from those of bacteria. (41).

In addition, we noticed that the C domains in fungus are significantly shorter than these in bacteria in all intermotifs, except C6-C7 (Fig S20C). In total, the whole region starting from motif C1 and ending in motif A1 is shorter by 13 aa in fungi by the median (Fig S20C). Considering that most of the fungal C domains are DCL subtype while bacterial C domains are mainly in LCL subtype, we also confirmed this difference in the DCL subtype, where the C1-A1 region is shorter by 19 aa in fungi (Fig S20D).

In summary, bacterial and fungal NRPS show significant difference in the length of intermotif. It makes it possible to distinguish between bacterial and fungal NRPS in our standardized architecture. And again, it suggests that bacterial and fungal NRPS have differentiated to different branches and should be treated differently.

Discussion

About twenty years ago, based on the first few crystal structures and a handful of sequences, the core motifs of NRPSs were mapped out (52). Since then, modules, domains, and motifs were used extensively utilized by annotation algorithms like antiSMASH(31) and Pfam(34) as well as experimental investigations (32,33), though with differing standards. Nowadays, with the explosive expansion of microbial sequencing data, systematic characterizations of NPRS is becoming possible. Here, we presented a motif-and-intermotif architecture of NRPSs by partitioning the C, A, T, and TE domains by well-known core motifs. Despite that all 19 motifs were derived from prior information, a standardized architecture presented a "common language" for processing NRPS sequences from various sources, allowing statistical insights from a large number of sequences without the need for manual curation.

Evolution imprints protein functions into their sequences. With the help of such standardized architecture, novel insights acquired from sequence statistics exhibit amazing concordance with our point mutation experiments, as well as earlier findings derived from structure or reengineering efforts. For example, we identified novel conserved residues distant from known motifs, among which the G-motif seems to be the most intriguing. It centers an exceptionally conserved glycine, where a single point mutation can abolish the enzymes function as we described in Figure 3. Interestingly, by visual cues from protein structure, G-motif has been used as the cut point of A subdomain for successfully reengineering A domain (62). In 2015, Kries et al. proposed the A domain reengineering strategy by A subdomain swap, as they discovered that in the protein structure that the substrate binding region is a flavodoxin-like subdomain. This subdomain starts from the middle of A3 and A4, and ends from the middle of A and A6 (the ending is exactly the G-motif, VPIGAPI in the PheA). Recently, another research team Thong et al. also used this A subdomain swapping strategy by CRISPR-Cas9 and successfully re-engineered A domain substrate specificity even subdomains from a range of NRPS enzymes of diverse bacterial origins (63).

NRPS undergoes substantial conformational changes as it exerts its multistep synthetase activity(55). Identifying key residues associated with such a degree of structural flexibility is crucial for understanding NRPS functionality. In our analysis, three additional conserved motifs were identified. We hypothesized that the G-motif in the loop might be one of these key sites for structural flexibility, acting as a hinge when NRPS switches between distinct states. First, the number of hydrogen bonds around the conserved glycine changes substantially during three function states. Also, glycine is the smallest in the 22 proteinogenic amino acid(64) , allowing the maximal extends of structural flexibility. Additionally, it has been reported that G-X-Y or Y-X-G oligopeptides motif provide flexibility necessary for enzyme conformation change for catalysis, where X and Y are polar and non-polar residues, respectively (65). G-R/K-P in the G-motif fit this G-X-Y structure. Furthermore, several enzymes have demonstrated the important function of conserved glycine residues in structure flexibility. For example, conserved G76 contributes to active-

site loop flexibility in the pepsin(66), and conserved G316 and G324 are structural basis of efficient metal exchange in the cadmium carbonic anhydrase of marine diatoms(67). In addition, the importance of this highly conserved glycine was verified in the fungal FQC biosynthesis system. We noted that the large amino acid in 409 position is mostly collided with N397 (F493 in this position of MiBiG NRPS sequences, Fig 2B-D) (Fig 3B). This may explain why mutation into large amino acids in 409 position always reduce the FQC yield. Nevertheless, mutation into a small amino acid proline also strongly reduces the yield. It is possible that prolines side chain locks the dihedral angle Φ of the protein backbone at approximately −65° and causes significant conformational restriction(68). Our results implied that the flexibility of glycine in the G-motif is essential for conformational change during A domain functions, therefore conformational restrictions in the G-motif impede the proper function of A domain.

We also found other conserved residues that may of functional importance. Similar to the G-motif, the conserved Ala and Pro in the start of A domain (Aα1) may bridge the A1 and the A4 motif. The conserved Glu acid and Leu located to the first helix of T domain (Tα1) may contribute to the correct folding of the T domain (for example: E699, L703, W707 and L711 in the first helix) , similar to the three conserved Gly contribute correctly folding and function in green fluorescent protein-like fluorescent proteins(69). From a functional perspective, these residues take part in formation of the hydrophobic core(58,59). In addition, 4'-phosphopantetheine cofactor required by T domain function is located in the vicinity of residue Y748, L751 and F752(70) (numbered in LgrA. They are L65, L68 and F69 in original research for holo-TycC3-PCP). Besides, Y748 and F752 also are the part of hydrophobic interface which interacts with C domain in docking process (in LgrA. V2534 and F2538 in original research for PCP2-C3 didomain from fuscachelin)(71). In summary, the conserved residues we identified may contribute to the stability and flexibility of NRPS protein in large conformational change for product synthesis. In the future, it would be meaningful to experimentally explore these new motif candidates more throughout, to gain deeper insights into the conformation dynamics during NRPS functioning.

Another point of consistency between our sequence characterization and earlier reengineering studies is the SCA sectors that we identified. Among three sectors we analyzed, the cutting point from Bozhüyük et al's. 2019 work(24) locates right at the beginning of the red sector, which contains residues enriched in the binding pocket of the A domain. The cutting points of Calcott et al. and Bozhüyük et al.'s 2018 work(23,25) both locate to the end of the green sector, which primarily represents covariation mode in the C domain. Congruence between sector boundaries and successful reengineering points substantiated SCAs potential to discern the "unit of evolution."

Nonetheless, our analysis also dissected and reemphasized the complexities in reengineering NRPSs. For example, sectors obtained from SCA intersperse with each other, even share covariation residues. Therefore, there is no universal cutting point in a NRPS module that does not disturb any of the major sector. This observation is in line with difficulties in reengineering NRPSs(20). The success of Bozhüyük et al. and Calcott et

al.(23-25) may be partially attributed to the fact that their cutting points are right besides some of the major sectors, and they utilized sequences from NRPSs relatively close on phylogeny. In the future, it might be possible to search in phylum for NRPS systems with non-overlapped covariation sectors, and use this system as the “building blocks” that can be recombined with more freedom. As homologous recombination requires conserved sequence, the standardized architecture could aid in establishing the appropriate “cutting points” in such potential NRPS systems for reengineering.

The phylogenys entanglement with the substr ate-specific binding pocket adds another layer of intricacy in manipulating NPRSs. The contribution of the phylogeny sector to the substrate specificity suggest non-degeneracy of the sequence space in determining substrate: same substrate can be selected by distinctive binding-pocket sequences, which drift with phylogenies and may recombine with each other. Such non-degeneracy leads to the diversity of the 10-S code, causing troubles in the targeted design of the A domain specificity. Despite we demonstrated that knowing the length of the five loops in the pocket region reduces the 10-S code diversity, a mechanistic understanding of substrate selectivity is still needed to guide the rational design of the A domain. The architectural distinctions between fungus and bacterium A domains discussed in earlier section may shed light on the detailed mechanism of substrate selection and establish the framework for developing a substrate prediction system, which could be less reliant on experimentally validated substrates.

In terms of limitations, most of our results is based on computation. Despite the high degree of concordance of computational predictions with our point mutation experiments on the G-motif, as well as with previous researches, more comprehensive experiments might be developed in future to explore the roles of other conserved residues we identified. Additionally, given the diversity of NRPS, despite our use of the largest A domain database with known substrates, A domains selecting rare substrates may be unrepresented. Overall, our effort is a preliminary investigation into the possibilities of standardized architecture in modular enzymes. Much more discoveries could be achieved in the future with the rapid expansions of microbial sequencing data.

Method

Data acquisition

We downloaded MiBiG database version 2.0(29), then extracted sequence information from this database. In this research, only C/A/T/E/TE domains from NRPS clusters were considered. Finally, 326 NPRS sequences from 264 species were obtained, with 1864 A domains, 1765 C domains, 1803 T domains, 310 E domains, 280 TE domains. In total, there are 1161 C-A-T modules, 685 C-A-T-C modules and 245 C-A-T-E modules. Sequences with high identity were removed (Fig S16A-C).

A domains sequences with known substrate were gathered from SeMPI 2.0(33). Some of the A domains from SeMPI overlap with these in MiBiG (http://sempi.pharmazie.uni-freiburg.de/database/MiBiG). Sequences which had deleted in MiBiG database or UniPort(72) were removed from our dataset. Finally, we collected 2636 A domain sequences. Sequence details see Table S3. Pfam seed alignment were downloaded for comparison with AntiSMASH domains (PF00668 for C domain, PF00501 for A domain, PF00550 for T domain).

Length threshold for multi-domain NRPS form MiBiG

Considering NRPS sequence length distribution, we filtered sequences by the following thresholds. For C-A-T, sequence length is more than 1000 and less than 1300; for C-A-T-C, sequence length is more than 1500 and less than 1750; for C-A-T-E, sequence length is more than 1350 and less than 2350. In practice, 14 sequences in the C-A-T-C type are removed because their secondary C domain, rather than the first C domain, will aligned with the first C domain of other sequences. And a few sequences from C-A-T-E were removed because they are incomplete at start. Exactly, they were full of gaps at start of MSA, even after C1.

Multiple sequence alignment

All multiple sequence alignments in our work were constructed by Cluster Omega 1.2.4(73).

AntiSMASH annotation

Data obtained from MiBiG has already been annotated by Antismash 5.1(74).

Conserved motif detection for NRPS domain

As shown by Fig S21, first, manual curation of the conserved motifs from literatures were performed for all domains. That include 7 motifs for C, 10 motifs for A, 1 motif for T, 7 motifs for E, and 1 motif for TE(52). A table of these conserved motifs according to previous literatures were summarized in table S3.

Second, we obtained reference sequences for each type of NRPS domain. The reference sequences of A domain, T domain and TE domain were from surfactin A synthetase C which has resolved crystal structure (PDB: 2VSQ)(50). The reference sequence of E domain was retrieved from the initiation module of tyrocidine biosynthesis (PDB: 2XHG)(51). The reference sequences of C domain were categorized by subtypes: the L-

type, D-type, dual and starter C subtype were curated from the work of Rausch et al.(53). The reference sequence of the heterocyclization subtype were extracted from crystal structure (PDB: 5T3E)(54). For consistency, motifs with similar positions to motifs in other C-type structural domains were taken instead of the core motifs defined in the original article in the heterocyclization type C domain. The motif locations on all reference sequences above were manually identified, according results from step 1.

Third, for every query sequence, global alignment was performed against possible reference sequences. Regions on the query sequence that align with the conserved motifs in the reference sequence were recorded as “motifs”.

Calculation of mutual information

In generating the fourth panel of Fig 1a, mutual information between the residues and the chirality subtypes of C domain was calculated. The C domain subtype was defined by the AntiSMASH annotation, including “LCL”, “DCL”, “Dual”, “Starter”, and “Heterocyclization”. In the multi-alignment of C domains, the reduced Shannon entropy on the amino acid composition at position *i* given the subtype, was calculated as the mutual information about chirality at position *i.*

The forth panel of Fig. 1b was obtained by the calculating the mutual information between substrate specificity and amino acid composition at each position of the multi-alignment of A domains.

Statistical coupling analysis

The statistical coupling analysis was performed with the binary approximation method(46). The cleaned correlation matrix is obtained using the method described in the same article(46).

To investigate substrate-related positions, we take substrates as the last column of the multiple sequence alignment. Then we apply SCA to this MSA. The least frequent substrate was taken as background of this position.

Loop length clustering

Pairwise the sum of Euclidean distance of five loops between different A domain sequences were calculated for distance matrix construction. To avoid the impact of local maximum value to the heatmap, the upper-bond in distance was set to 12. Then hierarchical clustering analysis is applied to this distance matrix (Fig S18).

The position of motif boundaries for multi-sequence

The motif boundaries in figure 2-4 and figure S13-14 are the average of motif boundaries position in each single sequence. Actually, except C domain, motif boundaries positions are consistent in different sequence after MSA. But, for C domain, they have different lengths in definition and low sequence similarity (Table S4).

Sector identification in SCA

We use the same notation and significant threshold in Halabi et al(46). The bra-ket notation is such that $|k\rangle$ denotes the kth eigenvector and $\langle i|k\rangle$ the weight for

position *i* along eigenvector *k*. The threshold ε = 0.05 is chosen to separate significant weights along an eigenvector from statistical noise.
For C-A-T, green sector is defined as $\langle i|2\rangle > \varepsilon$; magenta sector is defined as $\langle i|2\rangle < \varepsilon$; red sector is defined as $\langle i|6\rangle < \varepsilon$.
For C-A-T-C, green sector is defined as $\langle i|2\rangle > \varepsilon$; magenta sector is defined as $\langle i|4\rangle > \varepsilon$; red sector is defined as $\langle i|2\rangle < \varepsilon$.
For C-A-T-E, green sector is defined as $\langle i|2\rangle < \varepsilon$; magenta sector is defined as $\langle i|3\rangle > \varepsilon$; red sector is defined as $\langle i|2\rangle > \varepsilon$.
For A domain, from bottom to top, sectors are defined as $\langle i|1\rangle > \varepsilon$, $\langle i|2\rangle > \varepsilon$, $\langle i|3\rangle > \varepsilon$, $\langle i|4\rangle > \varepsilon$, $\langle i|5\rangle > \varepsilon$ and $\langle i|6\rangle < \varepsilon$.

Loop length profile calculation method

$$\overline{L_{i,j}} = \frac{L_{i,j} - \min(L_i) + 1}{mean(L_i) - \min(L_i) + 1}$$

Related to Fig S18, $\overline{L_{i,j}}$ is normalized length of loop i in jth row. $L_{i,j}$ is actual length of loop i in jth row. L_i is a column vector containing all length of loop i.

The calculation of conservation in NRPS protein

For a quantitative perspective, we use ConSurf(75) and AACon(76) to calculate conservation of specific position in the NRPS protein. The results of two methods are similar in very conserved positions, but a little different in middle conserved positions. The results are displayed in Table S1.

Comparison NRPS domains between bacteria and fungi

First, we obtain Phylum information form species name by TaxonKit(77). Then sequences are classified into bacteria and fungi, and sequences that are neither bacterial nor fungal are neglected.
We used two-sample Kolmogorov-Smirnov test to calculate the significance of difference. In practice, we used function kstest2 in Matlab. The meaning of symbol for significance as described below. (ns: no significance, P-value ≥ 0.05; two asterisks: 0.01 ≤ P-value < 0.05; three asterisks: 0.01 ≤ P-value < 0.05; four asterisks: 0.01 ≤ P-value < 0.05)

Experimental Materials and General Methods

The plasmids and strains utilized in this work are listed in Table S5. The oligonucleotide primers synthesized by Shanghai Sango Biotech are given in Table S6. PCR reactions were carried out with FastPfu high-fidelity DNA polymerase (Transgene Biotech) on a T100TM Thermal cycler from Bio-Rad. Escherichia coli strain DH5α was propagated in LB medium with appropriate antibiotics for plasmid DNA, and plasmid DNA was prepared using the Plasmid Mini Kit I (Omega). Automated DNA sequencing was performed by Shanghai Sango Biotech. PCR screening for transformants were carried out using 2×GS Taq pcr mix (Genesand).
LC-MS analysis was performed on an Agilent HPLC 1200 series system equipped with a single quadrupole mass selective detector, an electrospray ionization (ESI) and an Agilent 1100LC MSD model G1946D mass spectrometer by using a Venusil XBP C18 column (3.0

x 50 mm, 3 mm, Bonna-Agela Technologies, China) and equipped with an electrospray ionization (ESI) source.

Gene cloning, plasmid construction, and genetic manipulations

(a) Creation of *fmqC* deletion strains in *A. fumigatus*:

For site-directed mutagenesis of G-motif in *A. fumigatus*, *fmqC* was fully deleted in Cea17.2 to increase the rate of homologous recombination and reduce the difficulty of screening transformants. The FmqC (accession: EDP49773.1) deletion strain was created in Cea17.2 by replacing the *fmqC* with hygromycin phosphotransferase gene (*hph*) using modified double joint PCR described previously(78) consisting of the following: 1 kb DNA fragment upstream of *fmqC*, a 1.9 kb DNA fragment of *hph*, and a 1 kb DNA fragment downstream of *fmqC*. Third round PCR product were purified for protoplast transformation (fragment concentration greater than 300ng/μL). Polyethylene glycol (PEG) based transformation of *A. fumigatus* was done as previously described(79). The fragment concentration was 3 μg per 100 μL of protoplasts. The mutants were confirmed by using diagnostic PCR and the correct transformants were used for subsequent analysis.

(b) Creation of *fmqC* mutation strain in *A. fumigatus*

Seven plasmids were generated, one which included a full-length copy of *fmqC* (pYJY25), and the others were mutated *fmqC* (pYJY26-31). The pYJY25 plasmid was assembled by amplifying 1 kb DNA fragment upstream of *fmqC*, *fmqC*, *A. fumigatus pyrG* as the selectable marker, and a 1 kb DNA fragment downstream of *fmqC*. The fragments were assembled into a full plasmid by homologous recombination using Express® MultiS One Step Cloning Kit (Vazyme). pYJY26-31 were assembled using the same fragments and method, except that the targeted mutations of 409Gly in G-motif was introduced by using primers containing the mutation and amplifying *fmqC* in two separate PCR reactions. After plasmid construction, the fused 8kb PCR fragments with the point mutation were purified and used for transformation to TYJY81. Transformants were confirmed by sequencings (data not shown) to obtain TYJY82-88 for the subsequent experiments.

Detection of secondary metabolites by LC-MS analysis

A. fumigatus strains were cultivated in glucose minimal medium (GMM)(80) culture medium at 28 °C for 3 days in the dark. The media and mycelia were extracted with ethyl acetate, and then evaporated under reduced pressure. The extract was dissolved in 1 mL methanol, and 5 μL of the solution was directly injected for LC-MS analysis. Water with 0.1% (v/v) formic acid and acetonitrile were used as mobile phase with a flow rate of 0.5 mL/min. For analysis of the extracts, a linear gradient of acetonitrile in water (10-45%, v/v) in 30 min was used and washed with 100% (v/v) acetonitrile for 5 min.

Data Availability

NRPS Motif Finder is an online platform for construction of NRPS motif-and-intermotif architecture (http://www.bdainformatics.org/page?type=NRPSMotifFinder, Fig S22). The source code is available in supplementary material.

Acknowledgements

This work was supported by grants from Peking-Tsinghua Center for Life Sciences. We thank Song Chen, and Haonan Zheng for helpful discussions. We also appreciated the tech team at the BDA Informatics Suite (www.bdainformatics.org) for the online platform.

Funding

This work was supported by the National Key Research and Development Program of China (No. 2021YFF1200500), and National Natural Science Foundation of China (No. 32071255); The Biological Resources Program, Chinese Academy of Sciences (KFJ-BRP-009-005), and Key Research Program of Frontier Sciences, Chinese Academy of Sciences (ZDBS-LY-SM016).

Author contributions

Ruolin He collected the data, performed the majority of computational analysis in this research, drafted the manuscript, and suggested mutation sites for the experiments. Jinyu Zhang was responsible for execution and analysis of the mutation experiments. Yuanzhe Shao wrote python scripts for online platform. Shaohua Gu offered insightful comments and assisted in revising the manuscript. Wen-Bing Yin oversaw the project, designed experiments and revised the manuscript. Zhiyuan Li conceptualized and oversaw the project, conducted the initial analysis of the NRPS architecture, and revised the manuscript.

Competing interests

The authors declare no competing interests.

Reference

1. Wang, H., Fewer, D.P., Holm, L., Rouhiainen, L. and Sivonen, K. (2014) Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proceedings of the National Academy of Sciences*, **111**, 9259-9264.
2. Tyc, O., Song, C., Dickschat, J.S., Vos, M. and Garbeva, P. (2017) The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends in Microbiology*, **25**, 280-292.
3. Wei, B., Du, A.Q., Zhou, Z.Y., Lai, C., Yu, W.C., Yu, J.B., Yu, Y.L., Chen, J.W., Zhang, H.W., Xu, X.W. *et al.* (2021) An atlas of bacterial secondary metabolite biosynthesis gene clusters. *Environ Microbiol*, **23**, 6981-6992.
4. Kautsar, S.A., Blin, K., Shaw, S., Weber, T. and Medema, M.H. (2020) BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Research*, **49**, D490-D497.
5. González, O., Ortíz-Castro, R., Díaz-Pérez, C., Díaz-Pérez, A.L., Magaña-Dueñas, V., López-Bucio, J. and Campos-García, J. (2017) Non-ribosomal Peptide Synthases from Pseudomonas aeruginosa Play a Role in Cyclodipeptide Biosynthesis, Quorum-Sensing Regulation, and Root Development in a Plant Host. *Microb Ecol*, **73**, 616-629.
6. Morgan, G.L., Kretsch, A.M., Santa Maria, K.C., Weeks, S.J. and Li, B. (2019) Specificity of Nonribosomal Peptide Synthetases in the Biosynthesis of the Pseudomonas virulence factor. *Biochemistry*, **58**, 5249-5254.
7. Thomas, M.D., Langston-Unkefer, P.J., Uchytil, T.F. and Durbin, R.D. (1983) Inhibition of Glutamine Synthetase from Pea by Tabtoxinine-β-lactam. *Plant Physiology*, **71**, 912-915.
8. Lee, B.-N., Kroken, S., Chou, D.Y.T., Robbertse, B., Yoder, O.C. and Turgeon, B.G. (2005) Functional Analysis of All Nonribosomal Peptide Synthetases in *Cochliobolus heterostrophus* Reveals a Factor, NPS6, Involved in Virulence and Resistance to Oxidative Stress. *Eukaryotic Cell*, **4**, 545-555.
9. Cociancich, S., Pesic, A., Petras, D., Uhlmann, S., Kretz, J., Schubert, V., Vieweg, L., Duplan, S., Marguerettaz, M., Noëll, J. *et al.* (2015) The gyrase inhibitor albicidin consists of p-aminobenzoic acids and cyanoalanine. *Nature Chemical Biology*, **11**, 195-197.
10. Bhatt, K., Machado, H., Osório, N.S., Sousa, J., Cardoso, F., Magalhães, C., Chen, B., Chen, M., Kim, J., Singh, A. *et al.* (2018) A Nonribosomal Peptide Synthase Gene Driving Virulence in Mycobacterium tuberculosis. *mSphere*, **3**.
11. Zan, J., Li, Z., Tianero, M.D., Davis, J., Hill, R.T. and Donia, M.S. (2019) A microbial factory for defensive kahalalides in a tripartite marine symbiosis. *Science*, **364**.
12. Visca, P., Imperi, F. and Lamont, I.L. (2007) Pyoverdine siderophores: from biogenesis to biosignificance. *Trends in Microbiology*, **15**, 22-30.
13. Süssmuth, R.D. and Mainz, A. (2017) Nonribosomal Peptide Synthesis—Principles and Prospects. *Angewandte Chemie International Edition*, **56**, 3770-3821.
14. Du, L., Sánchez, C., Chen, M., Edwards, D.J. and Shen, B. (2000) The biosynthetic gene cluster for the antitumor drug bleomycin from Streptomyces verticillus ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. *Chem Biol*, **7**, 623-642.

15. Dreyfuss, M., Härri, E., Hofmann, H., Kobel, H., Pache, W. and Tscherter, H. (1976) Cyclosporin A and C. *European journal of applied microbiology and biotechnology*, **3**, 125-133.
16. Hollstein, U. (1974) Actinomycin. Chemistry and mechanism of action. *Chemical Reviews*, **74**, 625-652.
17. Fleming, A. (1929) On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. *British journal of experimental pathology*, **10**, 226.
18. Queener, S.W. (1990) Molecular biology of penicillin and cephalosporin biosynthesis. *Antimicrob Agents Chemother*, **34**, 943-948.
19. Yim, G., Thaker, M.N., Koteva, K. and Wright, G. (2014) Glycopeptide antibiotic biosynthesis. *The Journal of Antibiotics*, **67**, 31-41.
20. Brown, A.S., Calcott, M.J., Owen, J.G. and Ackerley, D.F. (2018) Structural, functional and evolutionary perspectives on effective re-engineering of non-ribosomal peptide synthetase assembly lines. *Natural product reports*, **35**, 1210-1228.
21. Hacker, C., Cai, X., Kegler, C., Zhao, L., Weickhmann, A.K., Wurm, J.P., Bode, H.B. and Wöhnert, J. (2018) Structure-based redesign of docking domain interactions modulates the product spectrum of a rhabdopeptide-synthesizing NRPS. *Nature Communications*, **9**, 4366.
22. Li, Y., Zhang, W., Zhang, H., Tian, W., Wu, L., Wang, S., Zheng, M., Zhang, J., Sun, C., Deng, Z. *et al.* (2018) Structural Basis of a Broadly Selective Acyltransferase from the Polyketide Synthase of Splenocin. *Angewandte Chemie International Edition*, **57**, 5823-5827.
23. Bozhüyük, K.A.J., Fleischhacker, F., Linck, A., Wesche, F., Tietze, A., Niesert, C.-P. and Bode, H.B. (2018) De novo design and engineering of non-ribosomal peptide synthetases. *Nature Chemistry*, **10**, 275-281.
24. Bozhüyük, K.A.J., Linck, A., Tietze, A., Kranz, J., Wesche, F., Nowak, S., Fleischhacker, F., Shi, Y.-N., Grün, P. and Bode, H.B. (2019) Modification and de novo design of non-ribosomal peptide synthetases using specific assembly points within condensation domains. *Nature Chemistry*, **11**, 653-661.
25. Calcott, M.J., Owen, J.G. and Ackerley, D.F. (2020) Efficient rational modification of non-ribosomal peptides by adenylation domain substitution. *Nature Communications*, **11**, 4554.
26. Stachelhaus, T., Mootz, H.D. and Marahiel, M.A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry & Biology*, **6**, 493-505.
27. Eppelmann, K., Stachelhaus, T. and Marahiel, M.A. (2002) Exploitation of the Selectivity-Conferring Code of Nonribosomal Peptide Synthetases for the Rational Design of Novel Peptide Antibiotics. *Biochemistry*, **41**, 9718-9726.
28. Palaniappan, K., Chen, I.-M.A., Chu, K., Ratner, A., Seshadri, R., Kyrpides, N.C., Ivanova, N.N. and Mouncey, N.J. (2019) IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Research*, **48**, D422-D430.
29. Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V. *et al.* (2019) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research*,

867 **48**, D454-D458.
868 30. Medema, M.H. and Fischbach, M.A. (2015) Computational approaches to natural product
869 discovery. *Nature Chemical Biology*, **11**, 639-648.
870 31. Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema,
871 Marnix H. and Weber, T. (2021) antiSMASH 6.0: improving cluster detection and
872 comparison capabilities. *Nucleic Acids Research*, **49**, W29-W35.
873 32. Skinnider, M.A., Dejong, C.A., Rees, P.N., Johnston, C.W., Li, H., Webster, A.L., Wyatt, M.A.
874 and Magarvey, N.A. (2015) Genomes to natural products PRediction Informatics for
875 Secondary Metabolomes (PRISM). *Nucleic Acids Res*, **43**, 9645-9662.
876 33. Zierep, P.F., Ceci, A.T., Dobrusin, I., Rockwell-Kollmann, S.C. and Günther, S. (2020) SeMPI
877 2.0-A Web Server for PKS and NRPS Predictions Combined with Metabolite Screening in
878 Natural Product Databases. *Metabolites*, **11**.
879 34. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, Gustavo A., Sonnhammer,
880 E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J. *et al.* (2020) Pfam: The protein
881 families database in 2021. *Nucleic Acids Research*, **49**, D412-D419.
882 35. Sonnhammer, E.L., Eddy, S.R. and Durbin, R. (1997) Pfam: a comprehensive database of
883 protein domain families based on seed alignments. *Proteins*, **28**, 405-420.
884 36. Rottig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C. and Kohlbacher, O. (2011)
885 NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity.
886 *Nucleic Acids Res*, **39**, W362-367.
887 37. Minowa, Y., Araki, M. and Kanehisa, M. (2007) Comprehensive analysis of distinctive
888 polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J*
889 *Mol Biol*, **368**, 1500-1517.
890 38. Khayatt, B.I., Overmars, L., Siezen, R.J. and Francke, C. (2013) Classification of the
891 adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of
892 substrate specific hidden Markov models. *PLoS One*, **8**, e62136.
893 39. Prieto, C., Garcia-Estrada, C., Lorenzana, D. and Martin, J.F. (2012) NRPSsp: non-ribosomal
894 peptide synthase substrate predictor. *Bioinformatics*, **28**, 426-427.
895 40. Challis, G.L., Ravel, J. and Townsend, C.A. (2000) Predictive, structure-based model of
896 amino acid recognition by nonribosomal peptide synthetase adenylation domains.
897 *Chemistry & Biology*, **7**, 211-224.
898 41. Lee, T.V., Johnson, L.J., Johnson, R.D., Koulman, A., Lane, G.A., Lott, J.S. and Arcus, V.L.
899 (2010) Structure of a Eukaryotic Nonribosomal Peptide Synthetase Adenylation Domain
900 That Activates a Large Hydroxamate Amino Acid in Siderophore Biosynthesis*. *Journal of*
901 *Biological Chemistry*, **285**, 2415-2427.
902 42. de Juan, D., Pazos, F. and Valencia, A. (2013) Emerging methods in protein co-evolution.
903 *Nature Reviews Genetics*, **14**, 249-261.
904 43. Korber, B.T., Farber, R.M., Wolpert, D.H. and Lapedes, A.S. (1993) Covariation of mutations
905 in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information
906 theoretic analysis. *Proc Natl Acad Sci U S A*, **90**, 7176-7180.
907 44. Sułkowska, J.I., Morcos, F., Weigt, M., Hwa, T. and Onuchic, J.N. (2012) Genomics-aided
908 structure prediction. *Proc Natl Acad Sci U S A*, **109**, 10340-10345.
909 45. Jones, D.T., Buchan, D.W., Cozzetto, D. and Pontil, M. (2012) PSICOV: precise structural
910 contact prediction using sparse inverse covariance estimation on large multiple sequence

alignments. *Bioinformatics*, **28**, 184-190.
46. Halabi, N., Rivoire, O., Leibler, S. and Ranganathan, R. (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell*, **138**, 774-786.
47. Smock, R.G., Rivoire, O., Russ, W.P., Swain, J.F., Leibler, S., Ranganathan, R. and Gierasch, L.M. (2010) An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol Syst Biol*, **6**, 414.
48. Reynolds, K.A., Russ, W.P., Socolich, M. and Ranganathan, R. (2013) In Keating, A. E. (ed.), *Methods in Enzymology*. Academic Press, Vol. 523, pp. 213-235.
49. Conti, E., Stachelhaus, T., Marahiel, M.A. and Brick, P. (1997) Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J*, **16**, 4174-4183.
50. Tanovic, A., Samel, S.A., Essen, L.O. and Marahiel, M.A. (2008) Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science*, **321**, 659-663.
51. Samel, S.A., Czodrowski, P. and Essen, L.O. (2014) Structure of the epimerization domain of tyrocidine synthetase A. *Acta Crystallogr D Biol Crystallogr*, **70**, 1442-1452.
52. Konz, D. and Marahiel, M.A. (1999) How do peptide synthetases generate structural diversity? *Chem Biol*, **6**, R39-R48.
53. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. and Huson, D.H. (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evolutionary Biology*, **7**, 78.
54. Bloudoff, K., Fage, C.D., Marahiel, M.A. and Schmeing, T.M. (2017) Structural and mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis. *Proc Natl Acad Sci U S A*, **114**, 95-100.
55. Reimer, J.M., Eivaskhani, M., Harb, I., Guarné, A., Weigt, M. and Schmeing, T.M. (2019) Structures of a dimodular nonribosomal peptide synthetase reveal conformational flexibility. *Science*, **366**, eaaw4388.
56. Ames, B.D., Haynes, S.W., Gao, X., Evans, B.S., Kelleher, N.L., Tang, Y. and Walsh, C.T. (2011) Complexity Generation in Fungal Peptidyl Alkaloid Biosynthesis: Oxidation of Fumiquinazoline A to the Heptacyclic Hemiaminal Fumiquinazoline C by the Flavoenzyme Af12070 from Aspergillus fumigatus. *Biochemistry*, **50**, 8756-8769.
57. Lim, F.Y., Ames, B., Walsh, C.T. and Keller, N.P. (2014) Co-ordination between BrlA regulation and secretion of the oxidoreductase FmqD directs selective accumulation of fumiquinazoline C to conidial tissues in Aspergillus fumigatus. *Cell Microbiol*, **16**, 1267-1283.
58. Weber, T., Baumgartner, R., Renner, C., Marahiel, M.A. and Holak, T.A. (2000) Solution structure of PCP, a prototype for the peptidyl carrier domains of modular peptide synthetases. *Structure*, **8**, 407-418.
59. Izoré, T. and Cryle, M.J. (2018) The many faces and important roles of protein-protein interactions during non-ribosomal peptide synthesis. *Nat Prod Rep*, **35**, 1120-1139.
60. Tahlan, K. and Jensen, S.E. (2013) Origins of the β-lactam rings in natural products. *The Journal of Antibiotics*, **66**, 401-410.
61. KONG, K.-F., SCHNEPER, L. and MATHEE, K. (2010) Beta-lactam antibiotics: from antibiosis to resistance and bacteriology. *APMIS*, **118**, 1-36.
62. Kries, H., Niquille, D.L. and Hilvert, D. (2015) A Subdomain Swap Strategy for

Reengineering Nonribosomal Peptides. *Chemistry & Biology*, **22**, 640-648.

63. Thong, W.L., Zhang, Y., Zhuo, Y., Robins, K.J., Fyans, J.K., Herbert, A.J., Law, B.J.C. and Micklefield, J. (2021) Gene editing enables rapid engineering of complex antibiotic assembly lines. *Nature Communications*, **12**, 6872.

64. Lehninger, A.L., Nelson, D.L., Cox, M.M. and Cox, M.M. (2005) *Lehninger principles of biochemistry*. Macmillan.

65. Yan, B.X. and Sun, Y.Q. (1997) Glycine residues provide flexibility for enzyme active sites. *Journal of Biological Chemistry*, **272**, 3190-3194.

66. OKONIEWSKA, M., TANAKA, T. and YADA, R.Y. (2000) The pepsin residue glycine-76 contributes to active-site loop flexibility and participates in catalysis. *Biochemical Journal*, **349**, 169-177.

67. Xu, Y., Feng, L., Jeffrey, P.D., Shi, Y. and Morel, F.M.M. (2008) Structure and metal exchange in the cadmium carbonic anhydrase of marine diatoms. *Nature*, **452**, 56-61.

68. Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, **12**, 345-364.

69. Nwafor, J., Salguero, C., Welcome, F., Durmus, S., Glasser, R.N., Zimmer, M. and Schneider, T.L. (2021) Why Are Gly31, Gly33, and Gly35 Highly Conserved in All Fluorescent Proteins? *Biochemistry*.

70. Koglin, A., Mofid, M.R., Löhr, F., Schäfer, B., Rogov, V.V., Blum, M.-M., Mittag, T., Marahiel, M.A., Bernhard, F. and Dötsch, V. (2006) Conformational Switches Modulate Protein Interactions in Peptide Antibiotic Synthetases. *Science*, **312**, 273-276.

71. Izoré, T., Candace Ho, Y.T., Kaczmarski, J.A., Gavriilidou, A., Chow, K.H., Steer, D.L., Goode, R.J.A., Schittenhelm, R.B., Tailhades, J., Tosin, M. *et al.* (2021) Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity. *Nature Communications*, **12**, 2511.

72. Consortium, T.U. (2020) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480-D489.

73. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*, **7**, 539.

74. Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H. and Weber, T. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*, **47**, W81-W87.

75. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T. and Ben-Tal, N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, **44**, W344-W350.

76. Agnieszka Golicz, P.V.T., Fábio Madeira, David M. A. Martin, James B. Procter and Geoffrey J. (2018).

77. Shen, W. and Ren, H. (2021) TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics*, **48**, 844-850.

78. Szewczyk, E., Nayak, T., Oakley, C.E., Edgerton, H., Xiong, Y., Taheri-Talesh, N., Osmani, S.A. and Oakley, B.R. (2006) Fusion PCR and gene targeting in Aspergillus nidulans. *Nature Protocols*, **1**, 3111-3120.

79. Bok, J.W. and Keller, N.P. (2004) LaeA, a regulator of secondary metabolism in Aspergillus spp. *Eukaryotic Cell*, **3**, 527-535.
80. Baccile, J.A., Spraker, J.E., Le, H.H., Brandenburger, E., Gomez, C., Bok, J.W., Macheleidt, J., Brakhage, A.A., Hoffmeister, D., Keller, N.P. *et al.* (2016) Plant-like biosynthesis of isoquinoline alkaloids in Aspergillus fumigatus. *Nature chemical biology*, **12**, 419-424.
81. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J.E. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, **10**, 845-858.

Figure and Table

Figure 1

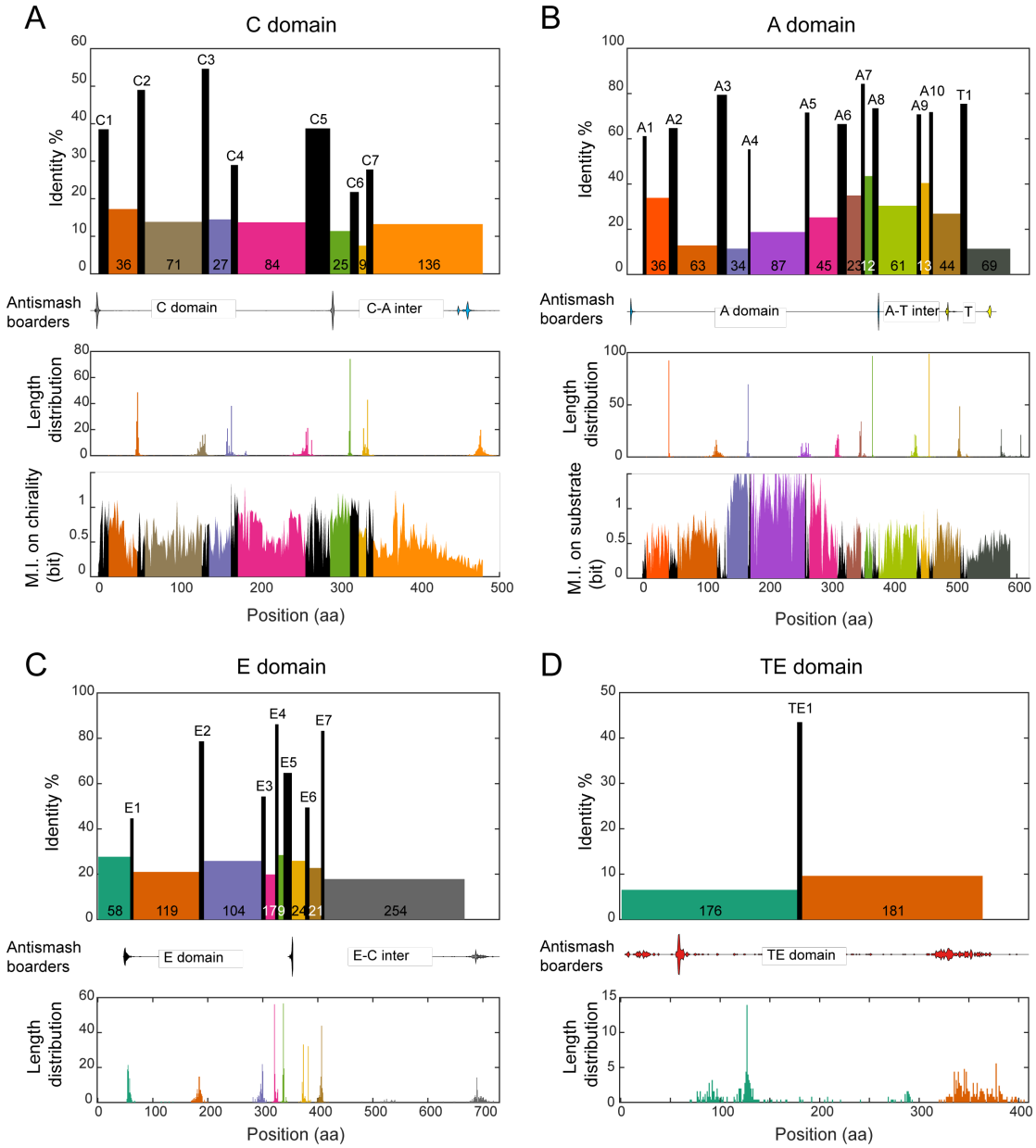


Figure 1. Overview of the motif-and-intermotif architecture of canonical NRPS domains.

A. Organization of C domain. Rectangles in the first panel show the seven C motifs (black) and eight inter-motif regions (colored), including the last one as the inter-motif region between C7 and A1 motif. The heights of the rectangles indicate the average sequence identities. The widths of the rectangles are their average lengths in amino acid sequences, with the average lengths of inter-motif regions printed on their rectangles. Violin plots in the second panel mark the border positions of domains annotated by Antismash 5.1 by the default option (the start and end borders of the C domain: gray; the start of the A domain: blue). The third panel shows the distributions

of the ending positions of the eight inter-motif regions, counting from their preceding conserved motifs. The fourth panel shows the mutual information (M.I.) between residues in the C domain and its chirality subtype.

B. Same as that in a, but for the organization of A domain and T domain. It starts with A1 motif and ends before C1 motif. And the fourth panel shows the mutual information between residues and A domain substrate specificity (see Methods for detail).

C. Same as that in A, but for the organization of E domain. It starts after T1 motif, and ends before C1 motif.

D. Same as that in C, but for the organization of TE domain. It starts after T1 motifs, and ends with regions after TE.

Figure 2

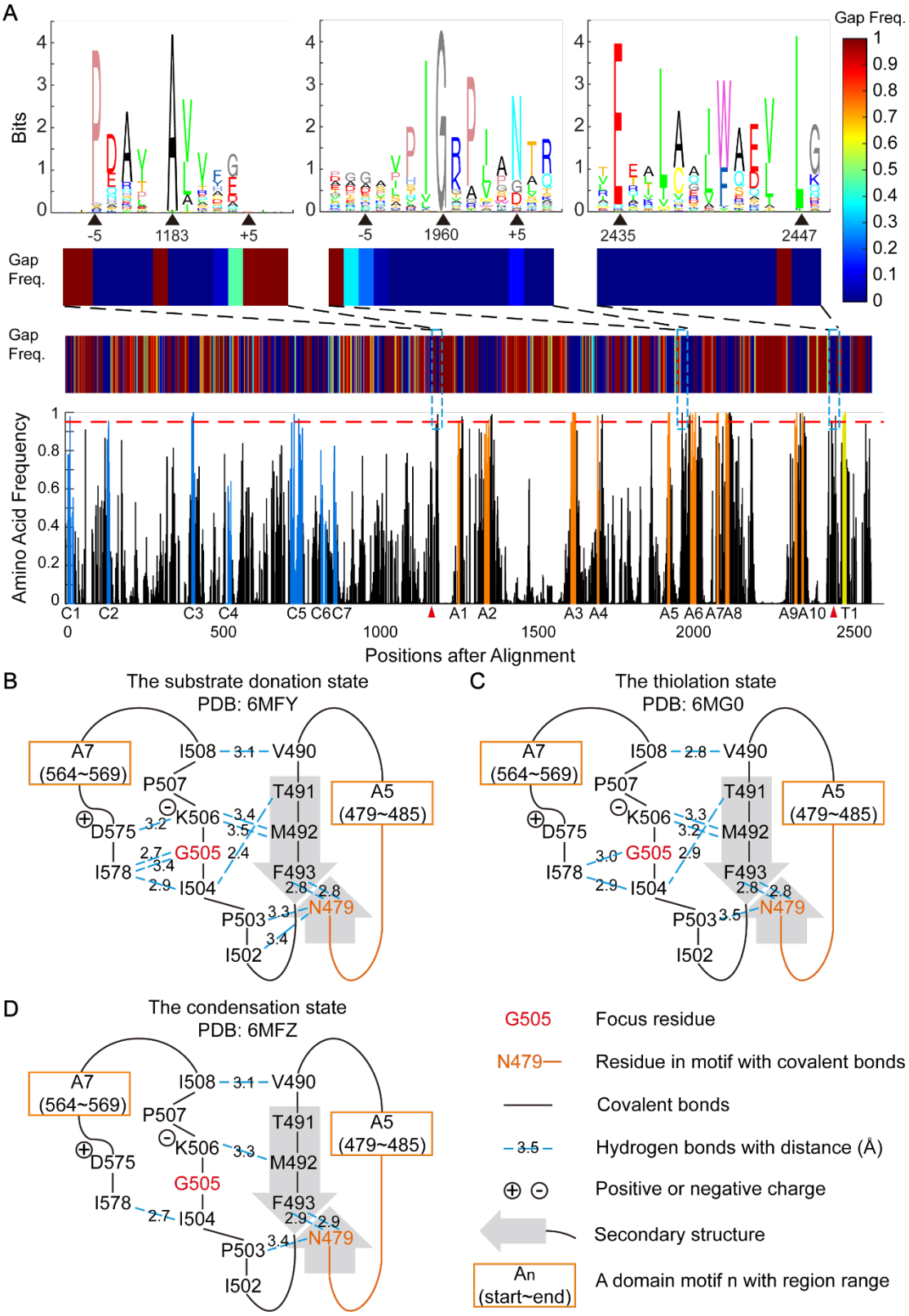


Figure 2. Analysis of amino acid frequency reveals potential new conserved motifs with implications in structural flexibility.

A. Amino acid frequency and gap frequency alone the multiple sequence alignment of NRPS CAT modules. In the bottom panel, bar heights indicate the frequency of the

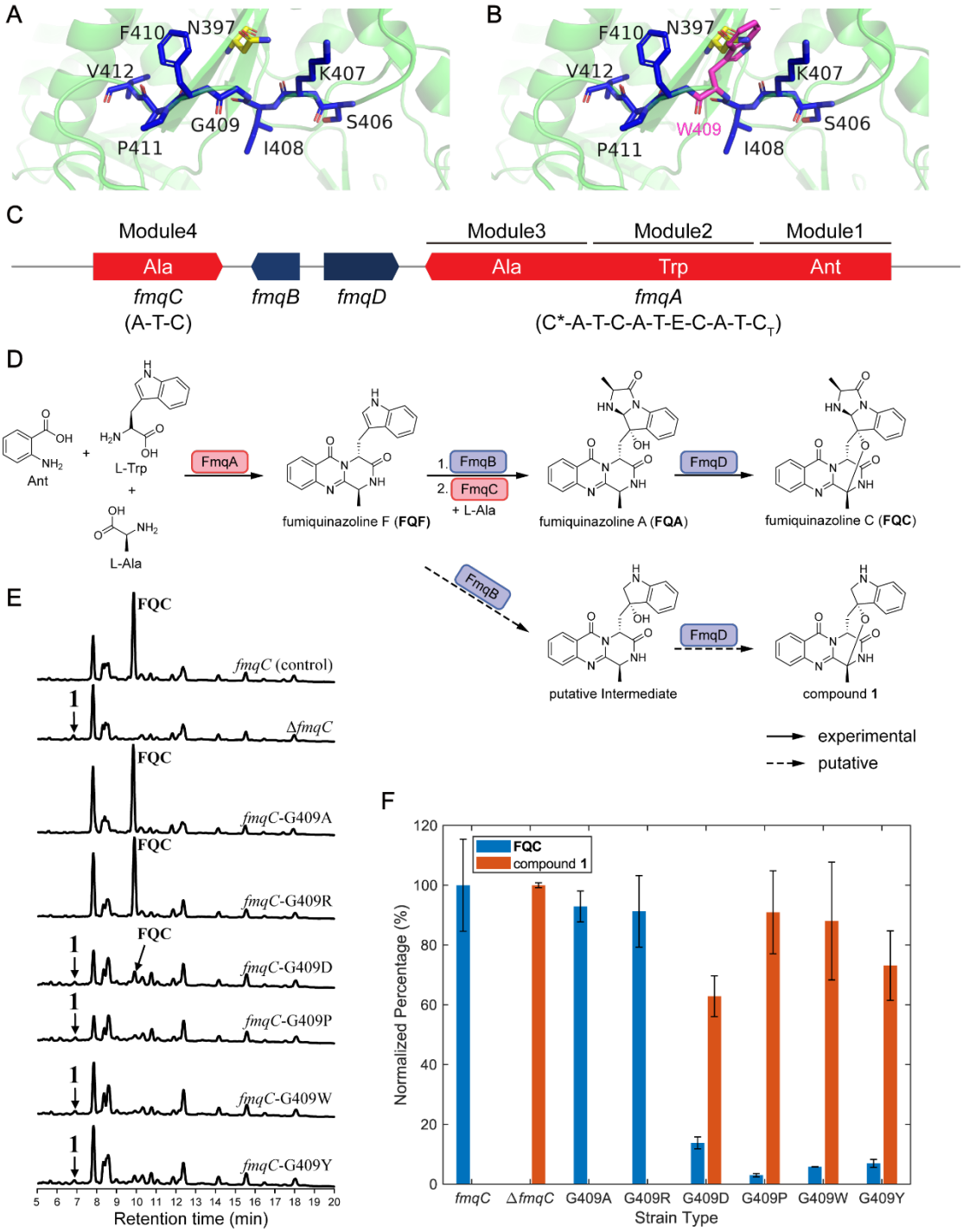
most frequent amino acid. Bars in the known core motifs from C, A, T domains were colored by blue, orange, and yellow, respectively. The horizontal red dashed line marks the frequent level of 0.95. Domain bondaries annotated by Pfam are divided by red triangles. The colored patch above amino acid frequency indicates the gap frequency, with blue color represents low gap frequency. Three potential new motifs (position 1183, 1960 and 2435/2447 in MSA) were marked by blue dashed box. The upper panel shows the sequence logo and the gap frequency near the three potential new motifs.

B. Chemical interactions and secondary structures surrounding the second potential new motif in (A) in the substrate donation state (PDB: 6MFY). Hydrogen bonds near the most conserved Gly were shown in blue dashed-line. Covalent bonds were shown as black lines. Secondary structures such as beta-sheets were shown as gray bold arrow. Known A domain motifs adjacent to related residues were shown in the orange box.

C. Same as that in B, but in the thiolation state (PDB: 6MG0).

D. Same as that in B, but in the condensation state (PDB: 6MFZ).

1071 Figure 3



1072
1073 **Figure 3. Mutations of G-motif G409 in FmqC supports the importance of the**
1074 **conserved domain in the biosynthesis of fumiquinazoline C**
1075 **A.** Residues in and near G-motif in the predicted structure of FmqC by Phyre2(81). G409
1076 is the conserved glycine in the G-motif. Residues in G-motif were marked by blue. The
1077 residue N397 (F493 in this position of MiBiG NRPS sequences, Fig 2B-D) which may
1078 collide which G409 was marked by yellow.
1079 **B.** Same as that in A, but with simulated mutation of G409W. The mutated tryptophan
1080 was marked in purple.

1081 **C.** The *fmq* gene cluster responsible for the production of FQC. Two NRPSs, *fmqA* and
1082 *fmqC*, were filled in red with their substrate selectivity marked. Ant: non-proteinogenic
1083 amino acid anthranilate. C* represents a truncated and presumably inactive C domain.
1084 C_T represents a terminal condensation-like domain which catalyzes macrocyclization
1085 reaction.
1086 **D.** The biosynthetic pathway for FQC is depicted, along with how it diverges into the
1087 production of compound **1** in the absence of functional FmqC.
1088 **E.** LC-MS analysis of the control *fmqC* (first row), Δ*fmqC* (second row) and six point
1089 mutation strains (3rd to 8th row).
1090 **F.** Normalized percentage of FQC and compound **1** in different strains. For FQC, the
1091 yield is normalized by its production in the wild type strain. For compound **1**, the yield is
1092 normalized by its production in the *fmqC* gene deletion strain. Error bars show standard
1093 deviations.
1094
1095

Figure 4

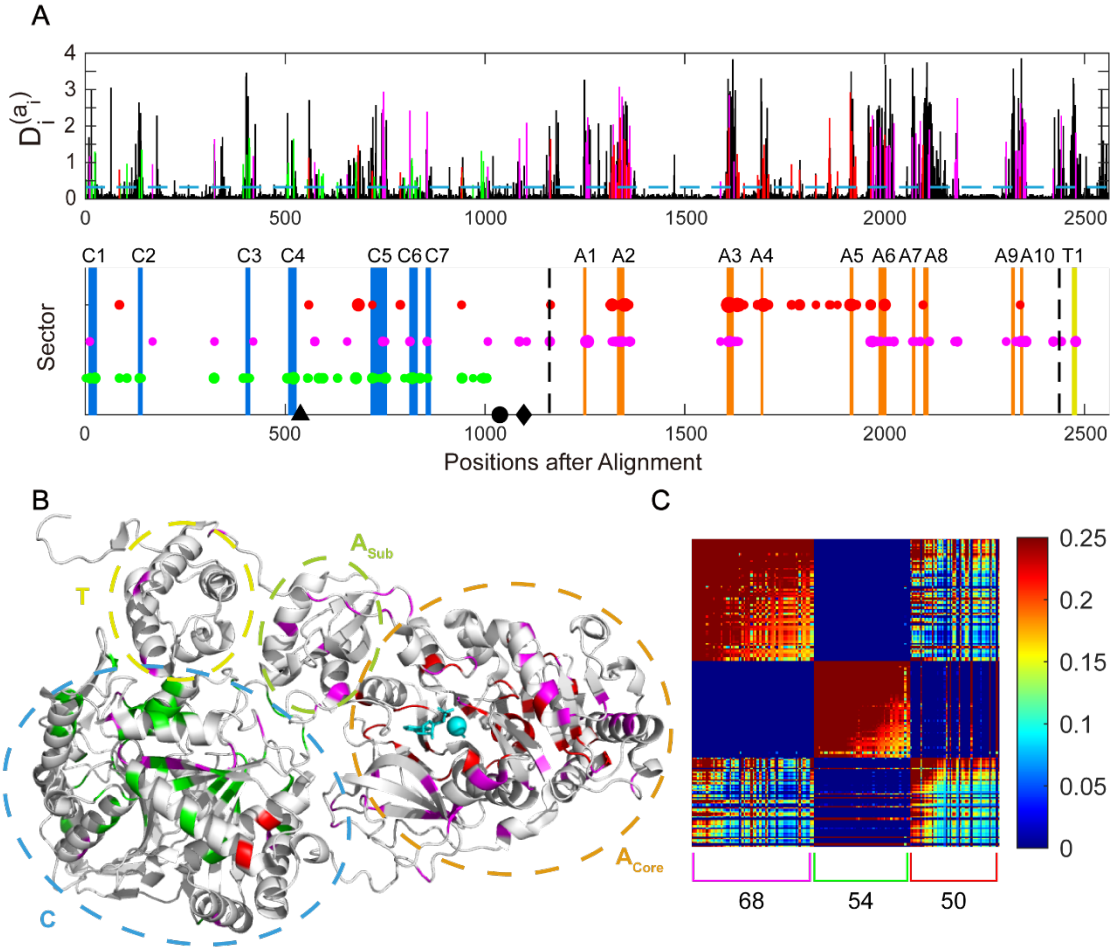


Figure 4. Statistical coupling analysis reveals overlapped sectors across C-A-T module

A. The upper panel shows the conservation of residues in a multiple sequence alignment of 1161 NRPS modules (containing C, A and T domain), quantified by the relative entropy $D_i^{(a_i)}$ in SCA method. Mean conservation level (0.32) is marked by the blue dashed line. In the lower panel, there are three groups of positions (green, magenta, and red, termed "sectors". Their corresponding conservations are marked in the same color in the upper panel. Blue bars mark C domain motifs from C1 to C7. Orange bars mark A domain motifs from A1 to A10. Yellow bar marks the T domain motif T1. Domain bondaries annotated by Pfam are divided by vertical black dashed lines. Black triangle marks reengineering point in the C domain reported by Bozhüyük et al.(24), black circle marks reengineering point in the C-A inter-domain reported by Calcott et al.(25) and black diamond marks reengineering point in the C-A inter-domain reported by Bozhüyük et al(23).

B. Mapping three groups of correlated conservation positions into the three-dimensional structure of NRPS module (PDB 4ZXI, containing C, A, T and Te domain. Te domain is hidden for clarity). Three sectors are marked in the same color as that in **(A).** C domain, A core domain, A sub domain, T domain are circled by blue, orange, limon, and yellow dotted line, respectively. Gly and AMP are substrates of this A domain. They and Mg2+ (for catalysis) are colored by cyan.

C. Heatmap of the SCA matrix after reduction of statistical noise and of global coherent correlations (See Method). Each sector is marked by corresponding color bracket under the heatmap, with the number of contained residues listed. 68, 54 and 50 positions belong to magenta green and red sector, respectively. In each sector, residues are ordered by descending contributions, showing that sector positions comprise a hierarchy of correlation strengths.

Figure 5

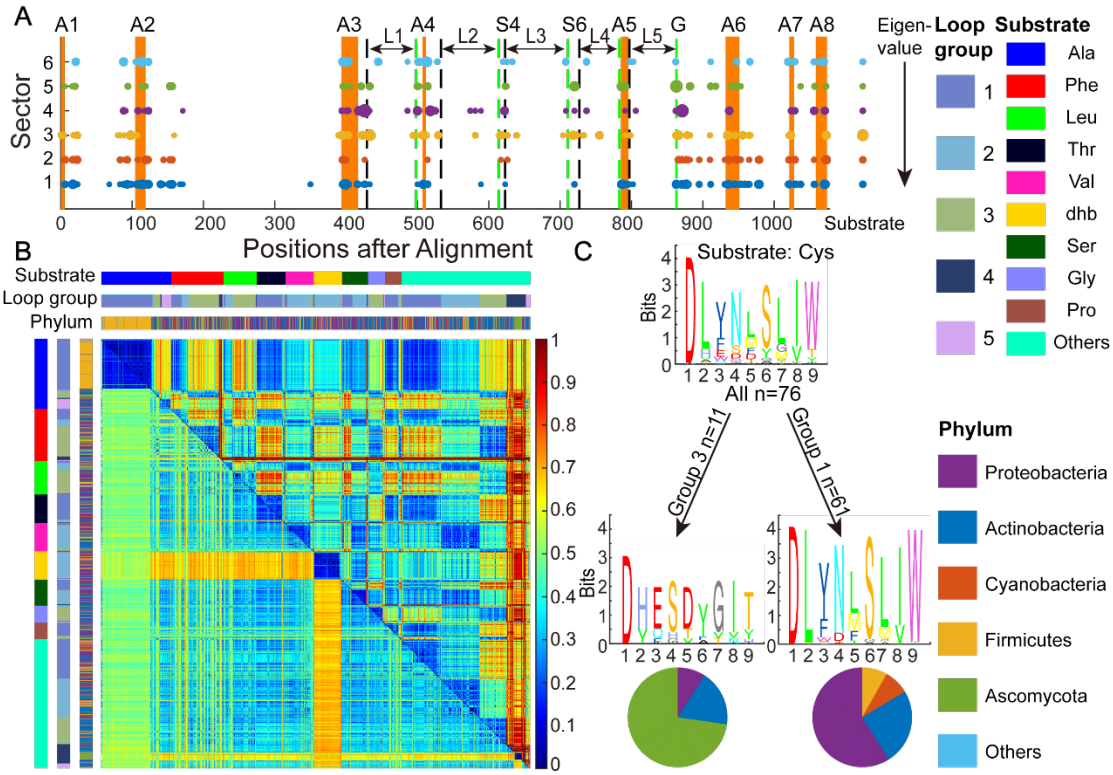


Figure 5. The specificity-conferring code of A domain is correlated with loop length and phylogeny.

A. SCA of 2636 A domain sequences, together with their substrate specificities attached to the last column of the multiple sequence alignment. Six sectors having a high contribution from the substrate column (>0.05, the size of points on the left scales the substrate's contribution to the sector, see Method) are sorted by their eigenvalues. The size of points scales its contribution to the sector. Orange bars mark A domain motif from A1 to A8. The start and end of five loop region are marked by black and green dotted line, respectively. S4 and S6 are the 4th and 6th of S codes. G is G-motif.

B. Distance matrix of A domain. Upper right heatmap is the Euclidean distance of the loop length as a 5-element vectors. Lower left heatmap is the sequence distance of the A domain. The matrix is sorted by the substrate specificity then the loop length group. Substrates, groups of loop length and phylum of these A domains are shown by colors in side bars.

C. Example showing that A domains conferring identical substrate exhibit distinct specificity-conferring codes, when they are categorized into different loop-length-groups. Phylum composition in each group is shown in the pie chart.

Figure 6

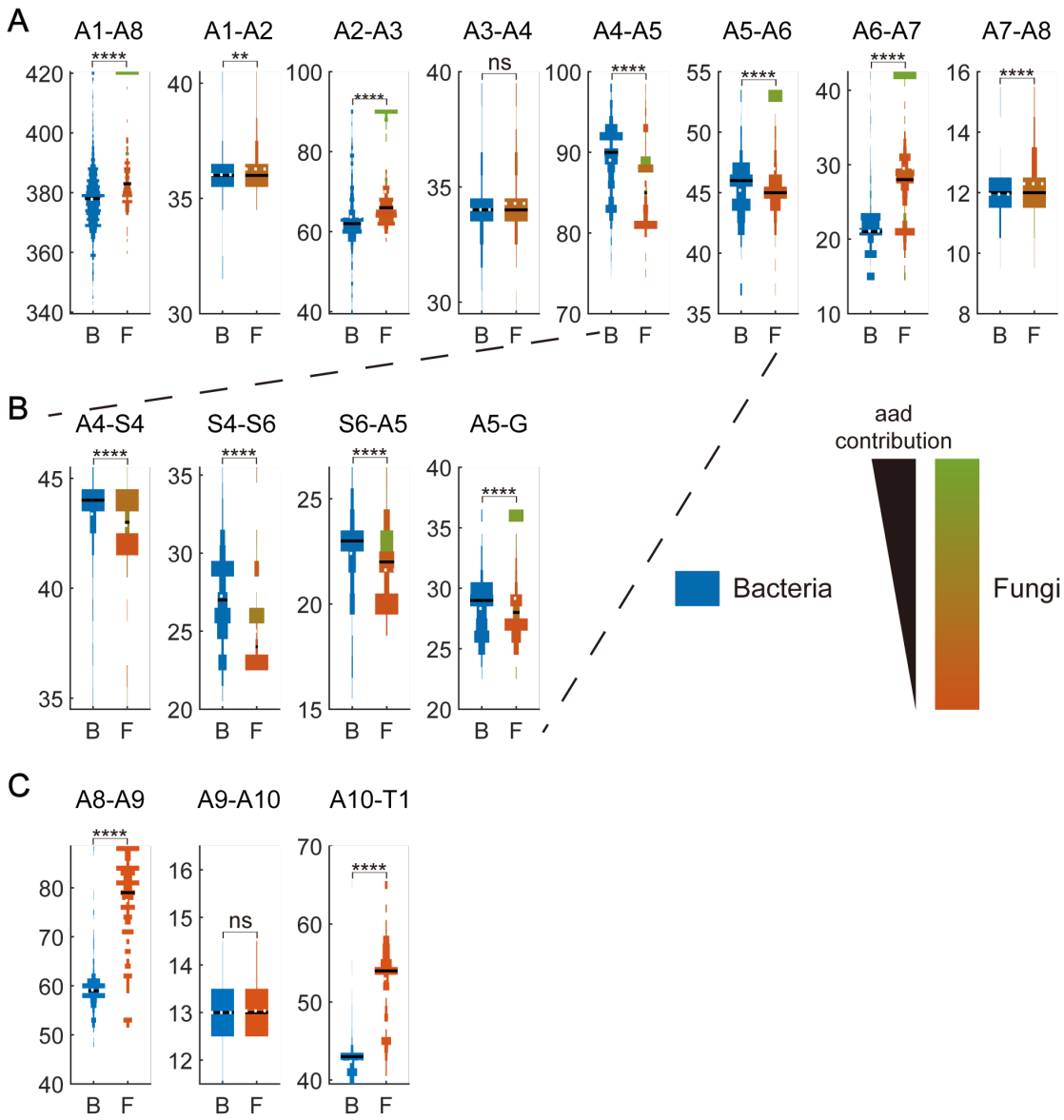


Figure 6. Comparison of NRPS architecture between bacteria and fungi in A domain

A. Length distributions of the inter-motif regions between motif Ai and Aj. Values of *i* and *j* are shown in titles. Blue color indicates distribution drawn from bacteria, while the green-orange color indicates distribution drawn from fungi. Color gradient bias towards green quantifies the contribution from Aad-specificity A domains to the histogram bin. Median values and mean values are marked by black lines and white dotted lines, respectively.
B. Same as (A), other than that it is for the loop length distribution in the A pocket.
C. Same as (A), other than that it is constructed by MiBiG database.

1165 Table1

The position of conserved residues		Amino acid composition	Properties of amino acid residues (side chains)
In the LgrA protein	**In the MiBiG C-A-T module MSA***	**in the MiBiG C-A-T module MSA**	
T491	T1927	T:61.67% S:15.16%	Polar, -OH; the donor and acceptor of hydrogen bonds
I504	I1959	I:88.29% V:5.25% L:3.45%	Hydrophobic
G505	G1960	G:99.66%	
K506	R1961	R:46.25% K:23.84%	Negatively charged; the donor of hydrogen bonds
I578	L2094	L:60.9% I:29.03% V:7.67%	Hydrophobic

1166 **Table 1. Representative chemically conserved residues in or interacting with the G-**
1167 **motif.**
1168 Conserved residues in or interacting with the G-motif have significant chemical properties.
1169 * This first letter is the prevalent amino acid in this position of MSA.
1170
1171