

Development and Application of Evolution and Big Data-Oriented Bioinformatics in Natural Product Research

Fanzhong Zhang^{1,2*}, Changjun Xiang^{1,2,3}, Lihan Zhang^{1,2}

¹Department of Chemistry, School of Science, Westlake University, Zhejiang Provincial Key Laboratory of Precision Synthesis of Functional Molecules, Hangzhou 310030, Zhejiang, China

²Institute of Science, Zhejiang Westlake Institute for Advanced Study, Hangzhou 310024, Zhejiang, China

³Department of Chemistry, Fudan University, Shanghai 200243, China

*Corresponding author: Fanzhong Zhang, zhangfanzhong@westlake.edu.cn

Copyright: © 2023 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract:

Billions of years of evolution in nature have nurtured abundant natural product resources, providing a vast molecular treasure trove for drug discovery and development. Evolution-oriented bioinformatics methods are playing an increasingly important role in the study of microbial natural products. The rapid growth of microbial genomic data presents new opportunities for big data analysis and evolutionary analysis of biosynthetic gene clusters. This not only allows us to have a clearer understanding of the panoramic view of natural products but also reveals the evolutionary patterns of natural products, utilizes evolutionary analysis methods and big data resources to discover novel drug lead natural products, understands biosynthetic enzymes, and even designs and modifies biosynthetic systems to create non-natural molecules. This article reviews recent advances in the application of evolution and big data-oriented bioinformatics to natural product research. It emphasizes the application of evolution and big data in the functional prediction of biosynthetic enzymes, evolutionary mechanisms, gene mining, and biosynthetic modification. Finally, it analyzes the current challenges and provides a view on future development trends.

Keywords:

Natural products
Evolution
Gene mining
Biosynthetic modification
Bioinformatics

Online publication: December 22, 2023

1. Introduction

Evolution is the process that promotes the emergence, development, and diversification of life^[1]. Essentially, the evolution of life is the evolution of genetic information. Enzymes involved in the biosynthesis of natural products

are encoded by genetic information, therefore, they are also subject to the forces of evolution^[1-3]. To adapt to natural environments, plants and microorganisms have created many natural products. Over the past century, natural products have played a significant role as lead

molecules in healthcare and agricultural production, including penicillin, erythromycin, and vancomycin, which have been used as drugs to benefit humanity. Bioinformatic predictive analysis indicates that only 3% of bacterial-derived natural products have been discovered so far, and even highly studied groups like *Streptomyces* still contain many unknown natural products^[4].

The development of gene sequencing technology has led to rapid growth in genomic data (**Figure 1**). The integration of large-scale genomics, metabolomics, and systematic data from functional studies, collectively known as “big data,” with bioinformatics, has brought a “technological revolution” to natural product research. Traditional natural product research heavily relied on compound isolation and purification, and natural products can only be understood through the accumulation of isolated monomeric compounds. Nowadays, it is transitioning to the stage of visualizing the panorama of natural products. Modern natural product research based on big data and bioinformatics has provided us with a macro-level understanding of the molecular diversity, abundance, and distribution of natural products. This allows us to appreciate the vast untapped potential of microbial natural product libraries and guides the discovery of new molecules with clinical or commercial value, enhancing the efficiency of natural product discovery^[1,2,5–8]. Due to the significantly lower number of sequenced fungal genomes compared to bacterial genomes (as of January 2023, the number of bacterial

genomes in the NCBI database was 1,420,776, while the number of fungal genomes was 28,183, accounting for only 2% of bacterial genomes), big data analysis mainly focuses on bacteria. Therefore, this article primarily elaborates on bacterial natural product research but also includes some fungal research.

Understanding the biosynthetic mechanisms of natural products and the biochemical characteristics of related enzymes has facilitated the application of evolutionary analysis in predicting enzyme functions, thereby guiding the modification of enzymes and biosynthetic pathways^[6,9]. Currently, research on natural products based on evolution focuses on the following aspects:

- (1) Discovering new natural products using evolutionary-guided methods (prediction of compound structures);
- (2) Predicting enzyme functions through evolutionary analysis;
- (3) Creating desired products by modifying biosynthetic systems.

Therefore, this article will focus on the progress in the application of evolutionary-guided bioinformatics methods based on big data in natural product discovery and enzyme engineering research, and provide insights into the development of these tools and methods in the field of natural product research.

2. Research strategies for natural products based on evolution and big data

Gene mining is the prediction and isolation of active natural products based on genetic information without prior knowledge of chemical structures^[10–13]. Microbial genome mining methods have revitalized antibiotic research, but these methods rely on sequence similarity searches of previously identified biosynthetic enzymes. This empirical nature limits the chemical space explored. In recent years, natural product researchers have incorporated evolutionary principles into genomic analysis to search for new pathways^[14,15]. Natural product research utilizing evolution and big data is based on predicting functional similarity through phylogenetic distance. When the target protein sequence is distant from and forms a different evolutionary branch from

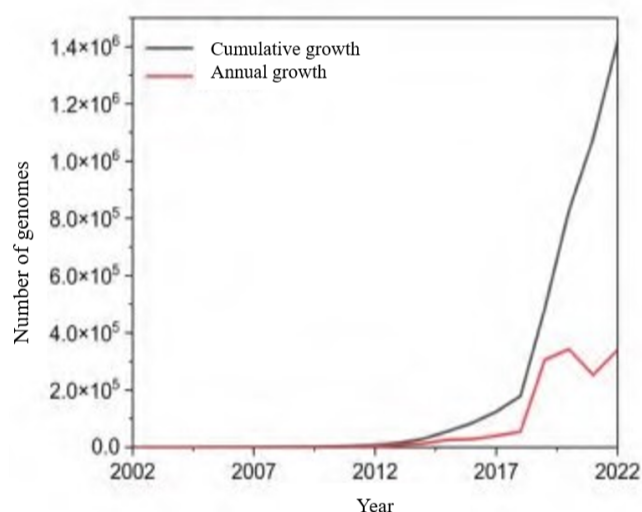
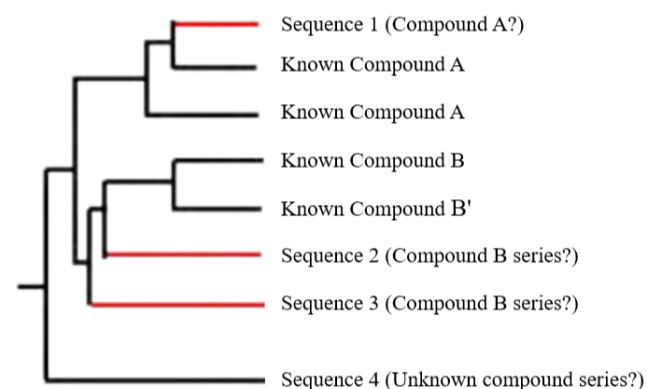
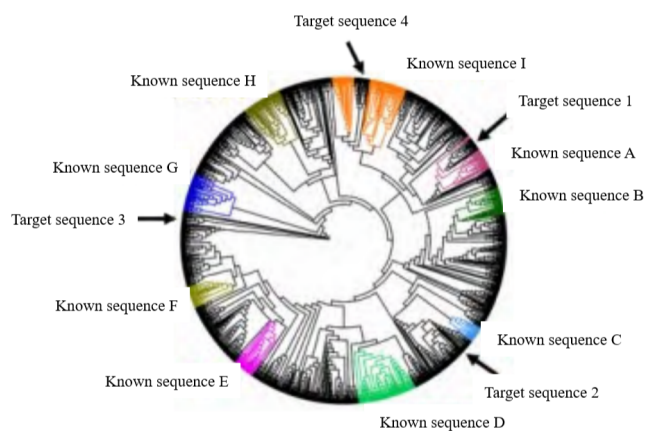


Figure 1. Growing trends for the number of bacterial genomes in the NCBI database from 2002 to 2022 (in the last two decades).

the encoding sequence of a known compound, it tends to produce a new product with a novel core structure. When the target sequence is adjacent to the encoding sequence of a known compound, it may produce a new product that is not significantly different from the known compound (**Figure 2a**)^[10,15]. Additionally, visualization of distribution patterns and diversity can be achieved based on whether the target sequence clusters with known sequences, belongs to a new branch, or is a rare outlier, presenting a panoramic view (**Figure 2b**).



(a) Prediction of functional similarity based on phylogenetic distance^[15].



(b) Schematic representation of the distribution patterns and diversity of target genes

Figure 2. An overall concept of the phylogenetic analysis.

There are several evolution-based bioinformatics tools available for natural product mining, such as ARTS (Antibiotic Resistant Target Seeker)^[16,17], NaPDoS (Natural Product Domain Seeker)/NaPDoS2^[18,19], EvoMining^[20], Big-SCAPE, and CORASON^[21,22]. ARTS and EvoMining are designed for evolutionarily related genomes, focusing on the prediction and cluster analysis of conserved biosynthetic gene clusters (BGCs).

ARTS targets resistance genes, automatically screens sequence data by linking housekeeping genes and known resistance genes to adjacent BGCs, mines antibiotics with novel modes of action, and compares similar BGCs and their putative resistance genes. EvoMining, based on gene duplication and substrate specificity expansion of enzymes, has developed a gene mining method that detects homologs of certain types of housekeeping genes and compares the average number and phylogenetic distance of enzyme families. It can intuitively showcase the origin and evolutionary direction of natural product biosynthetic enzymes. NaPDoS rapidly extracts and groups PCR products, genomic or metagenomic data, analyzes the position of target KS (ketosynthase) or C (condensation) domains on the evolutionary tree and infers the novelty and potential of secondary metabolites from bacterial genetic data. Big-SCAPE targets multiple genomes with unknown evolutionary information, using gene clusters from the MIBiG database^[23] as references to analyze gene clusters predicted by antiSMASH^[24]. It constructs a sequence similarity network, classifies these gene clusters into different gene cluster families, and then uses CORASON to interpret the evolutionary relationships among different gene clusters within each family. These bioinformatics tools are all based on evolutionary principles but target different types of genes and serve different purposes.

Currently, the best-studied class of natural product synthetic enzymes using evolutionary methods is modular enzymes, such as polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS). Based on the target gene cluster type, the study has classified the progress of evolution-guided research, mainly including PKS, NRPS, and other non-modular enzymes.

3. Evolution and big data-oriented PKS research

Polyketides are a large class of bioactive natural products with diverse structures and functions, and many clinically used drugs belong to this category, such as erythromycin, avermectin, and tetracycline. Polyketide biosynthesis is catalyzed by polyketide synthases (PKS). Currently, there are three known types of bacterial PKS (Type I, Type II, and Type III). Type I PKS forms an assembly

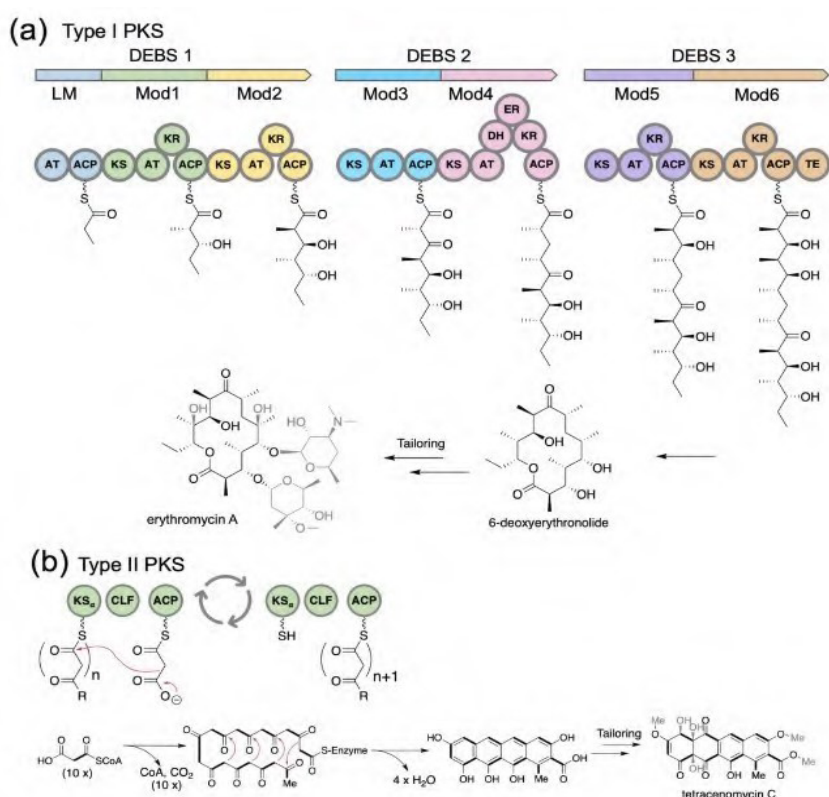
line composed of multiple modules, and each module contains core domains such as KS, AT (acyltransferase), and ACP (acyl carrier protein) that catalyze a cycle of polyketide chain extension. Some modules also contain domains like KR (ketoreductase), DH (dehydratase), and ER (enoylreductase) for varying degrees of polyketide modification ^[25] (**Figure 3a**). Type II PKS catalyzes the iterative condensation of acetate units through polyketide synthase (KS/KS α) and chain length factor (CLF/KS β), followed by reduction, cyclization, and dehydration reactions to form a polycyclic aromatic skeleton (**Figure 3b**) ^[26]. Type III PKS, also known as chalcone synthase-like PKS, belongs to the homodimeric enzyme and is essentially an iteratively acting condensing enzyme ^[27,28]. Many researchers have attempted to understand the relationship between PKS genes and polyketide structures in nature from an evolutionary perspective. This is primarily due to the potential discovery of new bioactive polyketides through exploring the diversity of natural PKS, the unique example provided by the multi-module structure of PKS for studying the evolution of multiple homologous but functionally distinct proteins, and the possibility of opening up new avenues for PKS

engineering through a better understanding of natural polyketide diversification mechanisms ^[29].

3.1. Evolutionary mechanisms of PKS genes

Through evolutionary analysis of different PKS domains (such as KS, AT, and KR), it is currently believed that the evolutionary processes leading to the diversification of PKS assembly lines mainly include gene duplication, horizontal gene transfer, gene conversion, and recombination (**Figure 4a**) ^[29]. The modularization of Type I PKS almost always originates from multiple copies of a single ancestral module ^[30]. Repeated modules provide an ideal platform for gene recombination, which can lead to corresponding changes in the chemical structure of the product ^[30]. Besides gene recombination, during gene evolution, DNA sequences may be non-interactively transferred from one homologous region to another, homogenizing these homologous sequences, a process known as gene conversion. Gene conversion is widespread in Type I PKS ^[31]. The analysis of these “natural reprogramming” events in PKS may aid in the development of biocombinatorial designs for bioactive compounds ^[30].

Figure 3. (a) Biosynthetic pathway of type I PKS (biosynthesis of erythromycin as an example); (b) Biosynthetic pathway of type II PKS (biosynthesis of tetracenomycin as an example).



related to the chemical structure of their substrates. Thus, they can be used for compound structure prediction and isomerase prediction. The KS domain sequence can not only distinguish between polyketides and fatty acids, enediynes, and polyunsaturated fatty acids but also between different types of polyketides, such as cis- and trans-AT PKS, PKS/NRPS hybrids. NaPDoS utilizes this principle for enzyme function and compound structure prediction^[18,19]. Additionally, the functions of modifying enzymes in type II PKS, such as the regioselectivity of KR and the cyclization mode of cyclases, can also be predicted through evolutionary analysis^[42,44].

Besides inferring the substrates and functions of similar enzymes in the same family based on enzymes with known functions, unknown enzymes can also be studied based on gene coevolution. BGC has undergone various evolutionary processes, such as intra-genomic duplication, rearrangement, domain/module/gene exchange, and horizontal gene transfer^[30,40]. Enzymes that interact within the same cluster require a coevolutionary process to maintain appropriate interactions^[6]. Therefore, the function of unknown enzymes can also be predicted based on the enzymes that interact with them. Crusemann's research group discovered a KS branch containing the TE B domain (responsible for O-acetylation) through evolutionary analysis of KS in trans-AT PKS. Based on the product structure and the missing HGTGT active site, it is inferred that these KS are non-extending KS0. Although they catalyze different polyketide structures, the TE B modules share biochemical consistency^[47].

In bacterial type I PKS, besides the KS domain, which can form evolutionary branches closely related to the chemical structure of the substrate, the AT domain also forms two main branches on the evolutionary tree, with specificity for receiving malonyl-CoA and methylmalonyl-CoA, respectively^[46]. The specific recognition of malonyl-CoA and methylmalonyl-CoA by the AT domain can be predicted through two characteristic regions in the sequence. The HAFH and GHS(I/V)G sequences indicate its reception of malonyl-CoA, while YASH and GHSQG indicate its reception of methylmalonyl-CoA^[47-49]. This discovery has long been used to distinguish the substrate selectivity of these ATs.

3.3. Gene mining of PKS

The theoretical foundation of PKS gene mining lies in the use of evolutionary analysis of the KS domain for compound structure prediction and isomerase prediction. Evolutionary analysis conducted on KS α and CLF (also known as KS β) within aromatic polyketide BGCs has revealed that the phylogenetic tree structure and branching patterns of KS α and CLF are highly similar, clustering based on the chain length of the corresponding polyketide compounds. Therefore, KS α and CLF can serve as ideal evolutionary markers representing the entire gene cluster^[42]. Brady's research group used CLF sequences as evolutionary markers to amplify related genes from soil microbiota. By comparing these sequences with known CLF genes and conducting evolutionary analysis, they discovered many sequences that fell into different sub-branches of the same clade as known sequences. Through heterologous expression of the corresponding gene clusters in the *Streptomyces albus* J1074 strain, they identified structurally novel and significantly active polyphenols and anthracycline compounds (1–3)^[50,51]. Recently, some research applied global genome mining in type II PKS and discovered oryzanaphthopyrans (4) from an evolutionary branch distant from known gene clusters. Evolutionary analysis based on big data also provided a comprehensive view of the distribution, abundance, and diversity of type II PKS^[5]. Li *et al.* (2022) utilized the evolutionary pattern of CLF combined with resistance gene targeting to mine tetracycline compounds, discovering the highly glycosylated tetracycline hainanmycin (5)^[52]. These studies, which either do not require microbial cultivation or can predict the structural novelty and bioactivity level of compounds before microbial cultivation, demonstrate the advantages of evolution-guided gene mining.

In type I PKS, the KS of trans-AT PKS can also form evolutionary branches closely related to the chemical structure of its substrates^[53]. Crusemann *et al.* utilized a KS database search to discover sequences closely related to mis PKS (misakinolides PKS) evolutionarily, ultimately identifying their product as the dimeric macrolide luminaolide B (6). By studying the biosynthesis and evolutionary relationships of misakinolides, scytophycin, and luminaolides, they found that their gene clusters originated from a common ancestor, achieving structural

diversification through the loss or acquisition of upstream or terminal PKS sequences^[54]. To enable structural prediction of trans-AT PKS products and understand the biosynthetic basis and evolutionary patterns of trans-AT PKS, Crusemann *et al.* and Medema *et al.* developed the online tools transATor and transPACT. TransATor takes PKS sequences as input and predicts KS substrate specificity, and the corresponding polyketide core structure^[55]. Using this tool, they discovered tartrolon-like compounds and leptolyngbyalide. TransPACT is a trans-AT PKS annotation and comparison tool that automatically forms functional branches of KS and identifies continuous modules shared by different PKS assembly chains. They utilized transPACT to obtain 1782 trans-AT PKS gene clusters from GenBank, analyzed them using antiSMASH, extracted KS sequences for evolutionary analysis, and performed gene mining based on the generated module-sharing network and phylogenetic tree. This led to the discovery of new trans-AT PKS products such as secimide (7), gynuellaide (8), and spliceostatin L (9), and explored the sequence-level correlations of similar chemical structures^[56]. These studies demonstrate that evolutionary analysis of the KS domain in trans-AT PKS can guide the discovery of structurally novel polyketide compounds, while also providing a foundation for PKS engineering to produce non-natural trans-AT PKS polyketide products.

Guo *et al.* (2016) utilized evolutionary analysis of KS sequences to mine type I PKS products from plant endophytic fungi, discovering the natural pigment talafun (10) with antibacterial activity. This study shows that using the highly conserved KS domain as an evolutionary marker can quickly link fungal genetic information and chemical structures, serving as a routine method for high-throughput sequencing technology in practical applications^[57].

Besides PKS itself, co-evolving genes within the same cluster can also be used for gene mining. Enediynes are a class of linear polyenes produced by type I PKS with extremely high activity, often used as antibody-drug conjugates in clinical trials^[58]. Shen *et al.* (2015) targeted two different enediyne biosynthetic gene sets, E5/E and E/E10, using real-time quantitative PCR to mine enediyne compounds from 3,400 strains. Through PCR, they identified 81 strains harboring enediyne polyketide synthase genes. Simultaneously, evolutionary analysis of gene E revealed that many clusters were distinct from known ones. To confirm these results, they performed genome sequencing on 31 representative strains, conducted GNN (Genome Neighborhood Network) analysis on the relevant gene clusters, and discovered gene clusters significantly different from known ones. Finally, through isolation and identification, they discovered the active compound tiancimycin A (11)^[58,59].

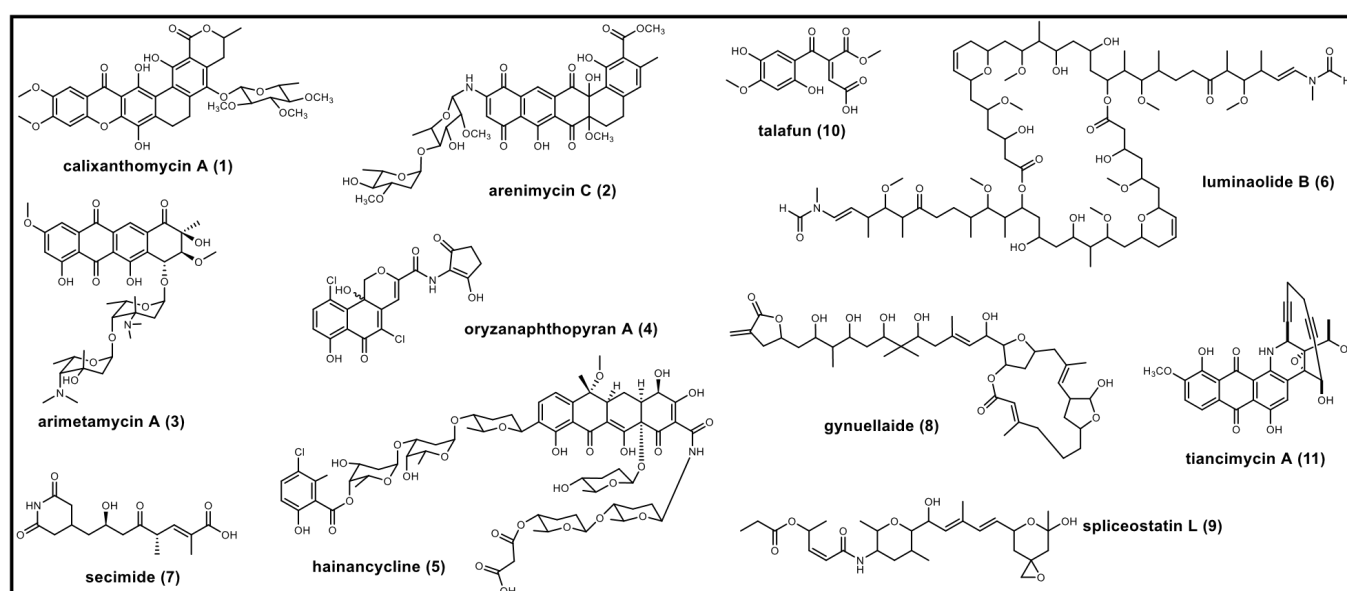


Figure 5. The structure of polyketides molecules 1–10 was obtained by phylogeny-guided genome mining (compounds 1–5 were aromatic polyketides discovered by genome mining of type II PKS. Compounds 6–9 were discovered by genome mining of trans-ATPKS. Compound 10 was discovered by genome mining of fungal type I PKS. Compounds 11 were enediynes).

These studies lay a foundation for mining more enediyne compounds or synthesizing enediyne homologs using PKS.

3.4. Biosynthetic engineering of PKS

In fact, since the identification of the modular characteristics of type I PKS more than thirty years ago, researchers have attempted to generate novel non-natural polyketide compounds through the recombination of modules and domains^[60]. Due to its modular nature, PKS provides a versatile synthetic platform, for example, as an effective method for synthesizing specific organic acids^[61]. However, in some early attempts to engineer the PKS assembly line, exchanging or deleting domains and modules often resulted in significantly reduced or even inactive enzyme activity^[52], presumably related to protein interactions^[53] and substrate selectivity^[54].

Simultaneously, increasing evidence suggests that a better understanding of the evolution of assembly line systems can further enhance the ability to engineer these systems (**Figure 6a**)^[29,35,62]. Drew *et al.* constructed multiple hybrid PKSs based on newly defined module boundaries, selecting cleavage points between KS and AT. The production of target compounds was significantly increased (10 to 48 times higher compared to hybrids constructed based on traditional definitions) (**Figure 6b**)^[63–65].

Hertweck *et al.* analyzed the phylogenetic trees of KS from various modules within several polyketide biosynthetic gene clusters. Through different methods of cleavage and fusion, they demonstrated that the addition or deletion of modules during natural evolution might occur at the KS-AT junction. By analyzing the substrate specificity of the P450 modifying enzyme within the gene cluster, they further inferred the evolutionary order of PKS^[66]. Similarly, besides the aforementioned KS-AT junction, the post-AT junction has also been proven to be an effective site for module fusion and domain exchange^[67–69]. A similar “cut-and-paste” strategy utilizing naturally preferred sites may also apply to the engineering of trans-AT PKS^[54].

Zargar *et al.* (2020) performed sequence alignments and exchanged KR or KR-DH-ER units between modules. Both *in vitro* and *in vivo* experiments demonstrated the feasibility of this strategy (**Figure 6c**)^[69–72], further indicating that the reducing domains may represent potential recombination units during evolution. On the other hand, some studies attempted to reverse the selectivity of AT^[73] and KR^[74] active sites through multiple point mutations. However, when these mutations were applied to the entire module, the expected product could not be obtained specifically. This suggests that PKS does not rely solely on point mutations but rather on

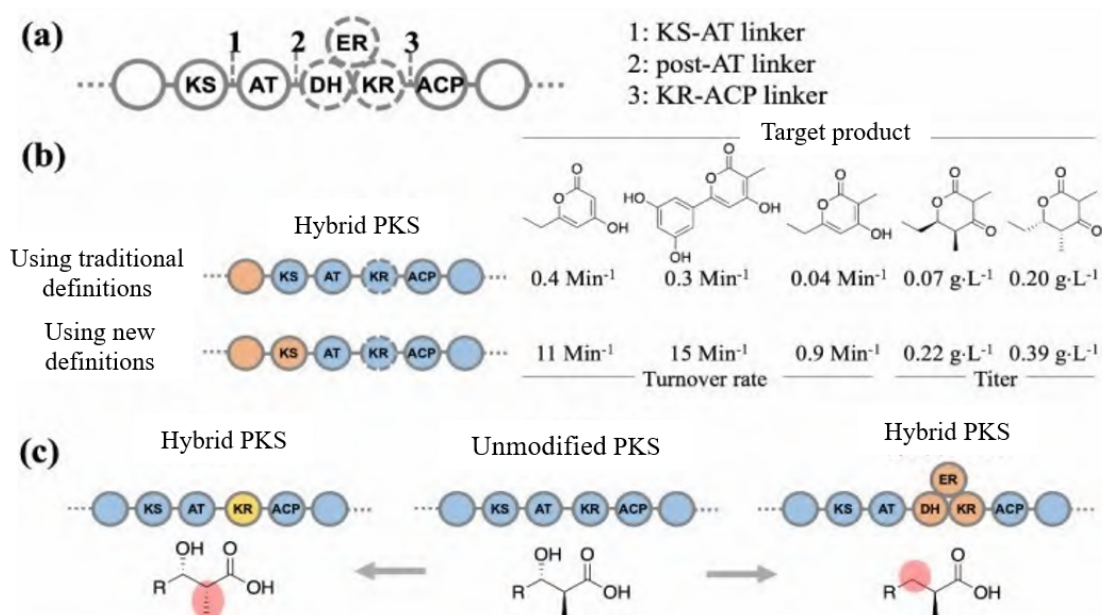


Figure 6. Evolution-guided engineering for PKS. (a) Natural recombination site for PKS; (b) Comparison of product yield between traditional definition-based and updated definition-based PKS engineering; (c) Engineering of reductive domains in PKS.

domain exchanges resulting from genetic recombination to alter domain selectivity^[29].

Although the structures of individual or multiple domains, and even entire modules of PKS, have been resolved through X-ray single crystal diffraction and cryo-electron microscopy techniques^[75–81], revealing the importance of protein interactions in various stages of polyketide chain elongation and some critical protein interaction sites, the underlying synergistic effects of various domains during the catalytic process of PKS have not been fully elucidated. Rational design and modification based on three-dimensional structures remain challenging. Therefore, analyzing the evolutionary relationships of entire PKS or individual domains from the perspective of natural evolution, inferring the sites of natural recombination as entry points for artificial modification, and selecting appropriate candidate PKS for hybrid PKS construction based on evolutionary relationships, provides researchers with a new evolution-guided approach for PKS modification.

4. Evolution and big data-oriented NRPS research

Non-ribosomal peptide synthetases (NRPSs) are multi-module enzymes or enzyme complexes derived from bacteria and fungi. Many of the peptide compounds they catalyze have important biological activities, and some are used clinically, such as cyclosporine, vancomycin, and daptomycin^[82–84]. Based on differences in their overall structures, NRPSs are typically classified into type I and type II^[85]. Type I NRPSs are large modular complexes that produce peptide compounds in a similar assembly line manner to type I PKSs. Each module mainly consists of three domains: C (condensation), A (adenylation), and T (thiolation, also known as a carrier protein), or other modifying domains such as E (epimerization). Type II NRPS proteins are typically independent enzymes or two domains that work together to form unique amino acid derivatives^[85]. During the synthesis of peptide compounds by NRPSs, the A domain selects specific amino acid monomers, which are activated by ATP to form aminoacyl-AMP and then transferred to the carrier protein T. The C domain condenses the activated aminoacyl (peptidyl) thioesters to extend the

chain through the formation of amide bonds. Like PKSs, NRPSs are significant for exploring active molecules and studying enzyme catalysis and protein interactions.

4.1. Evolutionary mechanisms of NRPS

Similar to PKSs, natural gene recombination plays a crucial role in the evolution of NRPSs. The diversification of non-ribosomal peptides is primarily driven by the recombination of A domains or subdomains^[86]. Recombination within the A domain occurs in the variable part of the A core to regulate substrates, while interactions between domains and the A sub are largely unaffected^[63,87].

Another core domain of NRPS, the C domain, can be classified into three types based on stereoselectivity: LCL, DCL, and starter C domain (CS or starter C). Although LCL and starter C domains show substrate differences (amino and β -hydroxy carboxylic acids) due to certain sequence variations, they appear to be more closely related to the evolutionary tree than other subtypes^[88]. Studies have indicated that the stereochemical selection of the C domain is related to the function of the E domain^[89]. In bacterial NRPSs, the PCPE-E-DCL sequence is almost universally conserved, suggesting that despite numerous genomic replication, insertion, deletion, and recombination events in evolutionary history, the E-DCL linker region has maintained strong selective pressure^[89].

4.2. Functional prediction of NRPS

Based on the evolutionary mechanism of NRPS, there is a direct relationship between the domain sequence on the NRPS assembly line and the product's chemical structure. This relationship makes it possible to predict the chemical structure of peptide compounds from DNA sequences.

In 1991, Stachelhaus *et al.* (1999) reported groundbreaking work on predicting the substrate specificity of A domains. They observed a strong correlation between the phylogenetic tree of A domains and substrate categories (**Figure 7a**). Secondly, they discovered that 10 key amino acid sequences (known as the Stachelhaus code) forming the substrate-binding pocket in the A core region are highly correlated with the substrates they receive^[90]. Subsequently, several tools for predicting NRPS A domain substrates have been

developed and utilized, such as NRPSpredictor2^[91] and SANDPUMA^[92]. The development of these prediction tools has greatly assisted in gene mining for discovering novel natural products, and on the other hand, they can also help find suitable candidate genes for NRPS assembly line modification.

The A domain of NRPS serves as an evolutionary signal at the “substrate level” and can be used to predict substrate specificity, while the C domain of NRPS acts as an evolutionary signal at the “pathway level” and can be employed to predict BGC patterns of similar molecules^[93]. Besides the original C domain, the C domain superfamily includes several other members that also belong to the NRPS domain, such as CS, DCL, LCL, E (epimerization), Cyc (heterocyclization), Dual C (epimerization/condensation), and modAA C (dehydroamino acid-related). Due to their different functions, these domains form distinct branches on the phylogenetic tree (**Figure 7b**)^[89,94]. Therefore, evolutionary analysis of the C domain can be utilized for functional prediction of related domains in NRPS.

4.3. NRPS gene mining

The substrate selectivity of the A and C domains in NRPS, along with the catalytic sequence of different modules, determines the order of amino acid connections. This means that the structure of non-ribosomal peptide compounds can be directly correlated with the NRPS sequence. Therefore, evolutionary analysis of the A or C domains can be utilized for NRPS gene mining.

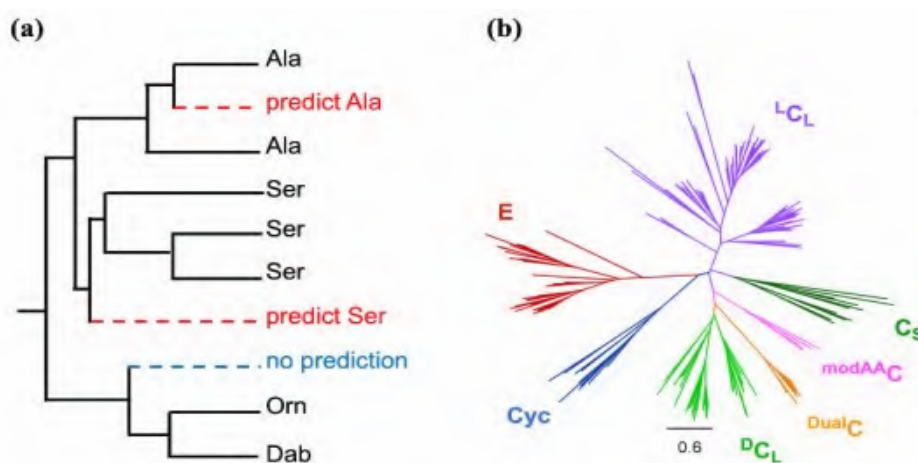
Calcium-dependent antibiotics are a class of cyclic

peptides that require calcium ions to exert their activity. Known compounds of this type share a conserved Asp4-X-Asp6-Gly7 segment that promotes their binding to calcium ions. Based on this, Brady *et al.* amplified the A domains of NRPS from soil eDNA using PCR and analyzed the amplified sequences using eSNaPD. The evolutionary tree of the tag sequences revealed many evolutionary branches of the Asp4 domain that are distant from known BGCs, suggesting the presence of unknown calcium-dependent antibiotics in the soil microbiome. Through heterologous expression, they identified a new class of calcium-dependent antibiotics called malacidins (11)^[95].

Although metagenome-based antibiotic discovery methods are still in their infancy, the scaled and automated approaches described in the above study provide a potentially powerful method for efficiently mining antibiotics hidden in the metagenome and combating antibiotic resistance.

In 2020, Culp *et al.* (2020) collected 71 gene clusters of glycopeptide antibiotics. Using the C domains from these gene clusters, they constructed an evolutionary tree to predict potentially new biologically active glycopeptide compounds. This led to the discovery of two novel glycopeptide antibiotics, corbomycin (12) and complestatin (13), which inhibit bacterial growth by binding to peptidoglycan and blocking the action of autolysins (peptidoglycan hydrolases necessary for cell wall remodeling during growth)^[96]. Recently, the same research group expanded their candidate BGCs using glycopeptide antibiotic fingerprint sequences.

Figure 7. (a) Prediction for substrates of A domains using phylogeny-guided method; (b) Unrooted phylogenetic tree of the C-domain superfamily^[90].



Through evolution-guided gene mining and heterologous expression, they identified five new type V glycopeptide antibiotics (type V GPAs), rimomycins (14), and misaugamycins (15). These antibiotics also inhibit cell division by preventing autolysin activity, demonstrating their mechanism of action^[97]. These discoveries expand the chemical diversity of type V GPAs, providing new chemical scaffolds for drug development and showcasing the significant potential of evolution-based bioinformatics platforms in mining the chemical “dark matter” of glycopeptide antibiotics.

Accurate prediction of NRPS has also facilitated the development of biologically active peptides independent of traditional isolation techniques. Brady *et al.* utilized bioinformatics predictions of lipopeptides and obtained a lipopeptide called cilagicin (16) through chemical synthesis, which exhibits strong antibacterial activity. Cilagicin exerts its antibacterial effect by blocking two essential undecaprenyl phosphates involved in cell wall biosynthesis^[98]. This study was based on an evolutionary tree analysis of the CS domain, leading to the discovery of an orphan BGC. The combination of compound structure prediction and chemical synthesis then yielded the corresponding product, circumventing issues such as non-expression of the target gene cluster or low product yield.

4.4. Biosynthetic engineering of NRPS and NRPS-PKS hybrids

4.4.1. NRPS

Currently, there have been many attempts to biosynthetically engineer NRPS, which can mainly be categorized into the following types: (1) Substituting A domains or A-T domains to change the extension units^[99]; (2) Modifying the substrate-binding pocket of A domains^[100,101]; (3) Swapping C-A or C-A-T domains^[102].

Through A domain substitution, Calcott *et al.*^[86] efficiently obtained high-yield modified pyoverdine peptides, determining the permissible boundaries for A domain recombination (**Figure 9a**). Crusemann *et al.*^[103] analyzed the nucleotide sequences of seven A domains in the hormaomycin biosynthetic gene cluster and found that, apart from approximately 400 base pairs related to the substrate recognition pocket, the remaining sequences showed over 90% similarity, suggesting potential natural recombination sites. Based on this hypothesis, using sequence boundaries inferred from natural recombination, three chimeras were constructed using the third A domain of HrmO as a template. *In vitro* experiments demonstrated successful transfer of A domain substrate specificity while maintaining a high conversion rate. Following the same strategy, Kries *et al.* transplanted nine different substrate specificities into the GrsA module of the gramicidin S

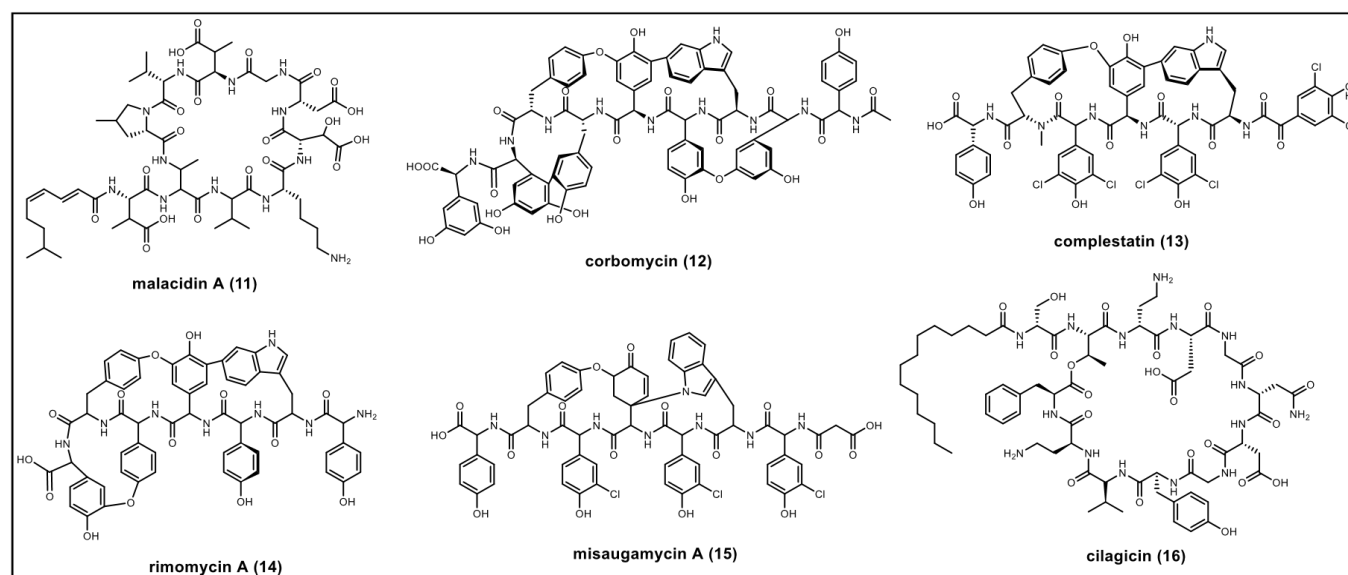
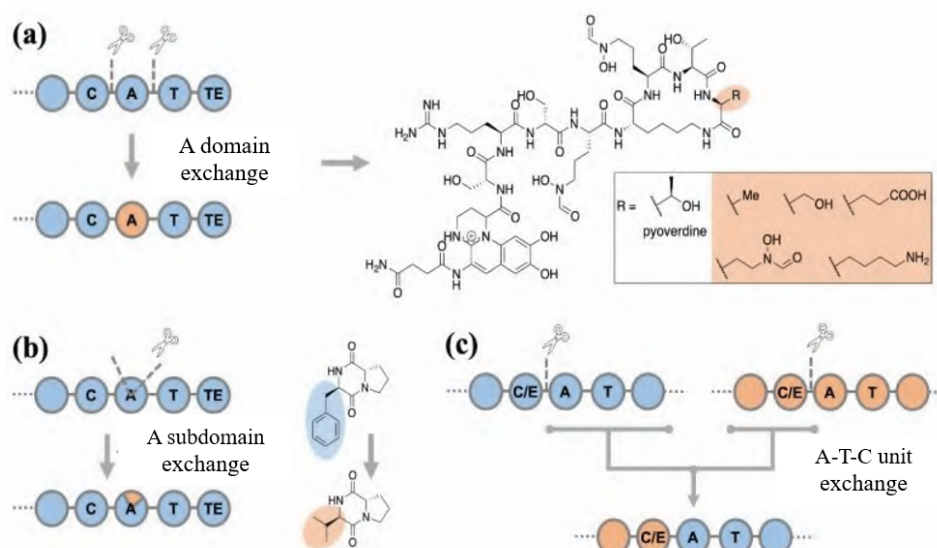


Figure 8. The structure of peptide molecules 11–16 was obtained by phylogeny-guided genome mining (compound 11 was a calcium-dependent antibiotic discovered by genome mining of A domain; compounds 12–15 belonged to the glycopeptide family of antibiotics discovered by genome mining of the C domain; compound 16 was a lipopeptide antibiotic discovered by NRPS prediction).

Figure 9. Evolution-guided engineering for NRPS. (a) Swap of A domain; (b) Swap of A subdomain; (c) Swap of A-T-C.



biosynthetic gene cluster (**Figure 9b**)^[104].

The C domain and C-A linker affect the catalytic activity and substrate selectivity of the A domain. Unlike strategies that only alter the core sequence of the A domain, Bozhüyük *et al.* (2019) defined A-T-C or A-T-C/E as exchange units, introducing the concepts of XU (exchange unit) and XUC (exchange unit condensation domain)^[105–108]. It was believed that the C-A linker is an ideal recombination site because sequence alignment shows low conservation in the A-T and T-C linkers, and they may be involved in important protein interactions during the catalytic cycle, while the interaction between C and A domains mainly relies on hydrophobic effects. Although hybrid NRPSs constructed based on these exchange units show reduced yields compared to the wild type, they can still produce enough target products for activity analysis (**Figure 9c**), and in a few cases, they achieve yields comparable to the wild type. Interestingly, this exchange unit is similar to the new definition of PKS module boundaries. Considering that the C domain also has a gating function, it can be inferred that there may be a certain correlation between the C domain and the upstream A-T domain in terms of evolutionary relationships.

4.4.2. NRPS/PKS hybrids

Polyketides and polypeptides have distinctly different backbones, and the hybridization of NRPS and PKS greatly promotes the diversification of natural product

types. Modifying NRPS/PKS assembly lines is an effective method to produce novel biologically active molecules. In fungi, highly reducing polyketide synthases (HR-PKS) can form hybrids with NRPS, synthesizing a series of fungal polyketide compounds represented by the pyridone skeleton. Minami *et al.* (2020) analyzed 884 PKS-NRPS hybrid enzymes in fungal genomes from NCBI and found a clear correspondence between the branches of the enzymatic phylogenetic tree and the molecular skeletons of the products, providing a macro perspective on the distribution and structural diversity of fungal PKS-NRPS gene cluster products^[109].

Bacterial PKS-NRPS hybrid enzymes are modular assembly line enzymes that synthesize compounds like bleomycin. Although many questions remain about the evolutionary mechanisms of PKS and NRPS^[29], evolutionary analysis has revealed some natural recombination sites that can serve as cutting points for module modification and domain fusion. The NRPS-PKS hybrid enzyme synthesizes the di-depsipeptide compound antimycins. Sequence analysis suggests that they may have evolved from the same ancestor as the trilactone JBIR-06 and the tetralactone neoantimycin A. Inspired by this, Ikuro *et al.* further inferred the site of natural recombination, added or subtracted modules from the JBIR-06 and neoantimycin A synthase, and achieved control over the ring size of the depsipeptide compounds through heterologous expression^[110]. This study confirms that the aforementioned PKS and NRPS engineering

strategies are also applicable to the NRPS-PKS hybrid enzyme system by analyzing the natural recombination and evolution process of NRPS-PKS hybrids ^[110].

5. Research on non-modular enzymes guided by evolution and big data

5.1. Mining of non-modular biosynthetic enzyme genes

5.1.1. RiPPs (Ribosomally-synthesized and post-translationally modified peptides) synthetase

Unlike polyketides and non-ribosomal peptides, the biosynthetic pathway of RiPPs lacks common biosynthetic features, making it difficult to perform reliable bioinformatics prediction on their gene clusters ^[111]. Prediction tools for RiPPs can rely on precursor peptide characteristics or modifying enzymes. Lu *et al.* (2022) explored the underlying logic of RiPPs biosynthesis through deep learning and proposed a combined model called BERiPPs (Bidirectional language model for Enhancing the performance of identification of RiPPs precursor peptides) based on the BERT pre-training model. BERiPPs can indiscriminately identify RiPPs precursor peptides without considering the genomic background and predict the cleavage site of the leader peptide, providing ideas for high-throughput mining of new RiPPs ^[112].

YcaO is a known modifying enzyme in RiPPs that catalyzes the formation of oxazolines, thiazoles, amidines, and thioamides. It can form thioamides when working together with the TfuA protein. Medema *et al.* (2013) screened 229 TfuA homolog proteins in actinomycetes and used a new gene mining tool, RiPPER, to retrieve their potential BGCs and 743 polypeptides, obtaining 74 different polypeptide networks. They then used MultiGeneBlast ^[113] to compare the gene clusters corresponding to each network and finally discovered a new class of thioamide compounds, thiovarsolins. They also demonstrated a strong correlation between TfuA evolution and precursor peptide similarity ^[111].

The biosynthesis of lasso peptides typically requires two enzymes: a lasso cyclase and a precursor peptidase. Tietz *et al.* (2017) developed an algorithm called RODEO for BGC identification, which outputs graphs and tables of BGCs and polypeptides for analysis. RODEO,

optimized for lasso peptides, summarized all potential lasso peptide gene clusters in GenBank, evaluated the obtained lasso peptides based on length and basic sequence features, and divided them into predicted leader and core regions. They constructed a sequence similarity network of 1315 lasso peptide precursors and identified six new lasso peptides ^[114]. This discovery expands the diversity of lasso peptides and provides a framework for future genomic mining of lasso peptides.

In addition, based on advances in machine learning technology, Merwin *et al.* (2020) developed DeepRiPP, which integrates genomic and metabolomic data and uses machine learning to automatically discover and isolate new RiPPs. DeepRiPP is implemented through three modules: identifying RiPPs independent of genomic structure and adjacent biosynthetic genes, preferentially selecting gene loci encoding new compounds, and automatically isolating corresponding products from complex bacterial extracts. They used DeepRiPP to perform large-scale comparative metabolomic analysis on a database of 10,498 extracts from 463 strains and finally discovered three novel RiPPs with structures fully consistent with platform predictions ^[115]. DeepRiPP improves the efficiency of RiPPs gene mining and demonstrates the application prospects of machine learning technology in microbial gene big data mining.

5.1.2. Terpene synthases

Terpenoids are important natural product types commonly found in fungi and plants. They are biosynthesized from linear precursors such as monoterpenes, sesquiterpenes, and diterpenes, formed by the condensation of IPP (isopentenyl diphosphate) and DMAPP (dimethylallyl diphosphate). Terpene synthases then catalyze diverse cyclization reactions to form complex carbon skeletons. Compared to plant and fungal terpene synthases, bacterial terpene synthases generally have low sequence similarity. Martin-Sanchez *et al.* (2019) conducted a whole-genome phylogenetic analysis of *Streptomyces*, comparing the distribution of terpene synthase genes and performing evolutionary analysis on these enzymes. They found that the evolution of these enzymes did not align with the evolution of *Streptomyces*, suggesting that horizontal gene transfer may be an important mechanism for the distribution of terpene synthase genes in *Streptomyces* ^[116]. Additionally,

they discovered that *Streptomyces* terpene synthases can be classified into ten groups on the evolutionary tree, with geosmin synthase being the most abundant. To explore the evolutionary relationship between bacterial and fungal terpene synthases, Avalos *et al.* (2022) conducted an evolutionary analysis of 908 fungal terpene synthases and 1,535 bacterial terpene synthases. Their study indicated that fungi also acquired terpene synthases from bacteria through horizontal gene transfer^[117]. Furthermore, there is increasing evidence in recent years that horizontal gene transfer plays a significant role in the evolution of terpene biosynthetic pathways^[118].

Diterpenes are terpenoids composed of four isoprene units, widely distributed in the plant kingdom. Many of their oxygenated derivatives, such as paclitaxel and triptolide, exhibit strong biological activities. Diterpenes have also been found in microbial metabolites, but compared to plant-derived diterpenes and enzymes, research on diterpene synthases from fungi is limited. To mine potential diterpene synthase encoding sequences from public databases, Yang *et al.* (2017) used the EriG protein (a cyclase forming the cyathane skeleton in *Hericium erinaceus*, belonging to the UbiA superfamily) sequence as a probe for genome mining. Through sequence clustering analysis and phylogenetic tree analysis, they discovered a new family of diterpene cyclases (cluster 11) related to UbiA in bacteria and fungi. By expressing and characterizing these enzymes in *Escherichia coli*, they identified seven new diterpene cyclases and determined the structures of their corresponding products, including a new diterpene called lydicene with an unusual skeleton^[119]. This study enriched the diversity of diterpene cyclases in bacteria and fungi, updated the members of the UbiA superfamily, and provided new opportunities for the application of microbial diterpene synthases in biocatalysis and metabolic engineering.

Recently, Chen *et al.* (2021) conducted a systematic evolutionary study on fungal sesquiterpene biosynthetic enzymes. Sesquiterpenes are synthesized by chimeric terpene synthases (PTTS) consisting of a C-terminal prenyltransferase (PT) domain and an N-terminal type I terpene synthase (TS) domain. Using the TS functional domains of 18 PTTSs, they constructed a phylogenetic tree and found that PTTSs form six major branches,

corresponding to different cyclization products^[120]. Chen *et al.* (2021) further expanded the PTTS phylogenetic tree analysis, combining gene mining to reveal that the six major branches roughly correspond to two major cyclization patterns of the isoprenoid linear precursor: the Type A reaction involving cyclization between the fourth and fifth double bonds, and the Type B reaction involving cyclization between the third and fourth double bonds^[120]. Utilizing PTTS phylogenetic tree-based gene mining, Tao *et al.* (2022) discovered three fungal-derived triterpene synthases and their corresponding triterpene products for the first time^[121]. Among them, the cyclization patterns catalyzed by triterpene synthases MpMS and CgCS do not fit into the aforementioned two categories, indicating the limitations of sequence-based functional prediction of terpene synthases and their rich catalytic plasticity.

5.1.3. Other genes

In PKS, the KS that catalyzes the Claisen condensation reaction to form the polyketide backbone belongs to the thiolase superfamily^[122]. Evolutionarily, members of this superfamily share functional clusters with branches and have diverged from a thiolase-like ancestor similar to archaeal thiolase. Considering their evolutionary diversity and structural similarity, Tan *et al.* (2020) hypothesized that enzymes other than PKS in the thiolase superfamily can catalyze iterative Claisen condensation reactions to synthesize polyketide skeletons. They demonstrated the feasibility of this pathway by synthesizing representative polyketide compounds such as lactones (triacetic acid lactones), alkylresorcinolic acids, alkylresorcinols, hydroxybenzoic acids, and alkylphenols^[123]. This discovery can be extended to other thiolases to further elucidate their structural and functional relationships and harness their biosynthetic potential for PKS research^[123].

Terminal alkynes are functional substances widely used in organic synthesis, medical science, materials science, and biochemistry. They can be catalyzed by a special desaturase enzyme, acetylene enzyme, in microorganisms^[124]. Zhu *et al.* (2015) elucidated the functions of JamA, JamB, and JamC in the biosynthesis of the terminal alkyne jamaicamide. Using the key enzyme for alkyne formation, jamB gene, as a probe, they performed evolutionary analysis on its sequence-similar genes to screen for new alkyne gene clusters. They

discovered a new terminal alkyne biosynthetic mechanism consisting of TtuA, B, and C, thereby expanding the research model for terminal alkyne biosynthesis and offering broad application prospects in synthesis and chemical biology^[125,126].

Indolocarbazole is a natural product used as a lead compound for anticancer drugs. Its core structure is formed by the dimerization of two molecules of oxidized tryptophan catalyzed by chromopyrrolic acid (CPA) synthase. Phylogenetic analysis of CPA synthase homologous genes in soil metagenomes led to the discovery of new indole tryptophans, borregomycins A-D, erdasporine A-B, and reductasporine^[15,127,128].

Ansamycins are an important family of natural products in clinical settings. These compounds are characterized by the presence of an aromatic core derived from the common precursor 3-amino-5-hydroxybenzoic acid (AHBA). Evolutionary analysis of AHBA synthase homologous genes in *Streptomyces* resulted in the discovery of 25 ansamycins from six strains, including eight new compounds such as juanlimycins and neoansamycin^[15].

Evolution-guided gene mining methods involve the evolutionary analysis of biosynthetic genes for compounds, such as acetylene enzymes that catalyze terminal alkyne formation, CPA synthases for indolocarbazole formation, and AHBA synthases for ansamycin formation. This approach allows for the discovery of active compounds with similar functional groups and enables the retrieval of special compound BGCs that cannot be directly detected using antiSMASH or ClusterFinder^[129]. It provides a feasibility validation for developing evolution-based bioinformatics approaches.

5.2. Natural product discovery targeting resistance genes

One of the primary goals of natural product discovery is to identify new antibiotics with novel modes of action to combat multidrug resistance in pathogenic bacteria. To avoid harming themselves with the antibiotics they produce, microorganisms have evolved several resistance strategies to circumvent self-toxicity. These strategies include product modification, substrate transport and binding, target duplication or modification, and are encoded by resistance genes located near the antibiotic

BGC^[130]. The presence of resistance genes within a gene cluster can serve as a window to predict the biological activity of the natural product synthesized by that pathway. Natural product discovery based on self-resistance gene identification helps bridge the gap between activity-guided and genome-guided methods in natural product discovery and functional assignment^[10,131].

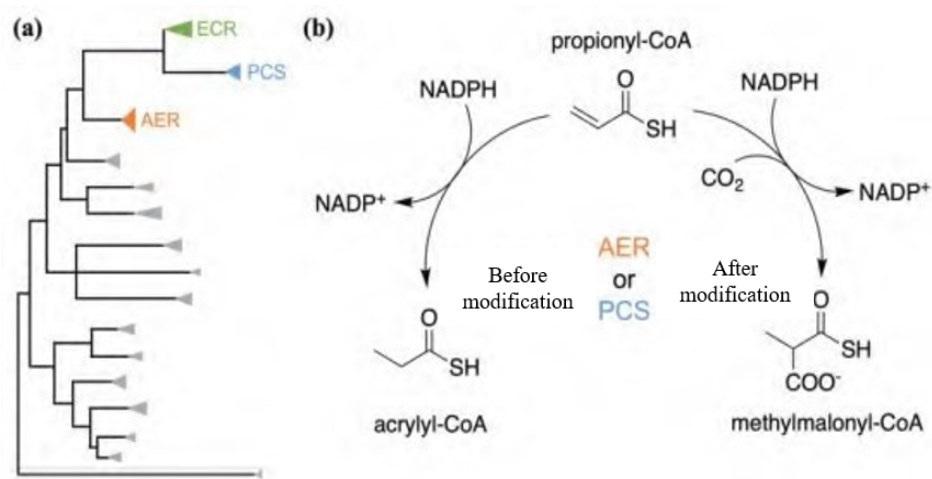
In recent years, there have been efforts to utilize evolutionary thinking in targeting resistance genes for natural product discovery. For example, the aforementioned ARTS is an evolution-guided gene mining tool that targets resistance genes^[16,17]. Using the ARTS detection mode, Adamek *et al.* analyzed all known BGCs and available bacterial genomes containing known drug-resistance target genes. Besides the 26 known gene clusters in the MIBiG database, they detected 22 additional gene clusters with resistance targets, demonstrating the potential of evolution-guided gene mining.

The transcriptional regulators tetR/marR and resistance transporters like tetA are a common pair of resistance genes in tetracycline biosynthesis. Resistance transporter proteins are tetracycline/metal proton antiporters located on the cell membrane, while the regulatory protein TetR is a tetracycline-inducible repressor. Based on the resistance mechanism of the tetracycline BGC, Li *et al.* (2022) used TetR/MarR-transporters as indicators for mining tetracycline-like natural products. Combining this with a phylogenetic analysis of chain length factors (CLFs) for further refinement, they discovered 25 different tetracycline gene clusters and ultimately isolated a new tetracycline called hainanmycin^[52]. This gene mining approach, targeting both resistance genes and type II PKS genes, offers the possibility of specific and efficient discovery of novel antibiotics.

5.3. Evolution-guided non-modular enzyme engineering

Biosynthetic enzyme engineering inspired by evolutionary analysis is not limited to multi-modular enzymes such as PKS and NRPS. Bernhardsgrutter *et al.* (2019)^[132] performed cluster analysis (**Figure 10a**) by combining enoyl-CoA carboxylase/reductase (ECR) with other medium-chain dehydrogenase/reductases (MDR).

Figure 10. (a) Phylogenetic analysis of MDR; (b) Engineering for AER and PCS.



They found that propionyl-CoA synthase (PCS) and archaeal enoyl reductase (AER) may have evolved from a common ancestor with ECR and possess potential CO₂ binding pockets and carboxylation functions. Under a certain CO₂ concentration, both PCS and AER exhibited weak carboxylation activity, but mainly produced reduction products (greater than 95%). Combining with a three-dimensional structural model, the authors further analyzed the sequences of the CO₂ binding pockets and mutated key amino acids. By enhancing CO₂ binding and preventing water from entering the pocket, they successfully activated the carboxylation functions of PCS and AER. The proportion of carboxylated products increased by about twenty times, becoming the main product (**Figure 10b**).

The aforementioned research elucidates the evolution of specific protein functions through the comparison of homologous proteins, a method commonly referred to as “horizontal.” This approach is based on the analysis of proteins found in extant species at a particular evolutionary stage. However, phylogenetic algorithms add a vertical dimension to sequence analysis, enabling the tracing of common ancestors from extant sequences^[133]. Ancestral sequence reconstruction (ASR) is a powerful tool for inferring original sequences from modern (i.e., extant) sequences^[134]. A fundamental element of ASR is the computation of phylogenetic trees, where the leaves represent selected extant sequences, and the reconstructed sequence associated with the tree root represents the common ancestor of the studied sequences. If this sequence encodes a protein, the ancestral protein can

be “resurrected” through gene synthesis techniques and studied for its biochemical properties using biochemical experiments. ASR also allows for the derivation of sequences for all internal nodes in the tree, further elucidating evolutionary processes^[133].

6. Summary

BGCs can effectively disperse through widespread horizontal gene transfer, even crossing the boundaries of phyla. Thus, evolutionary-based discovery strategies in natural product research powerfully complement traditional methods in terms of increasing novelty. The development and application of bioinformatics analysis tools based on evolutionary principles have generated growing genetic (gene), catalytic (protein), and chemical (compound structure) databases. These advances have propelled natural product research into the modern era of big data, making it possible to visualize the panoramic landscape of natural products^[1,2,6]. Natural products obtained through these strategies deepen the understanding of the synthetic pathways of naturally bioactive molecules and enrich the library of bioactive compounds.

As the quantity and quality of natural product-related research increase, the potential for applying artificial intelligence analysis methods, such as machine learning, is also growing. Combining evolution-guided approaches with artificial intelligence represents one of the future directions in this field. Successful machine learning methods require high-quality training data, which may necessitate coordinated efforts across laboratories and

even internationally to generate and manage datasets in a standardized manner^[135]. This relies on the development of bioinformatics^[136].

Natural product mining based on evolution and big data is inherently limited by the amount and scope of sequenced genomic data and faces several challenges:

- (1) Many microorganisms containing target BGCs are non-culturable under laboratory conditions or do not express the target gene clusters. Currently, the main solution to this problem relies on the development of molecular biology techniques such as heterologous expression. Therefore, improvements in techniques such as heterologous expression efficiency will facilitate evolution-guided natural product mining;
- (2) Predicting the biological activity of gene cluster products remains difficult. Currently, the discovery of active molecules can only be enhanced by targeting analogs of bioactive molecules or natural products with resistance genes. Exploring the relationship between compound structure and biological activity in evolution may provide more opportunities for evolution-guided natural product mining to discover new active molecules;
- (3) BGCs for terpenes, alkaloids, and other compounds often do not exhibit structural features of the compounds, and predicting the molecular structure of gene cluster products remains a significant challenge. For these non-modular BGCs, we need to increase the number of characterizations of gene clusters and their products and carefully analyze the evolutionary

characteristics and patterns of each biosynthetic enzyme step.

In terms of enzyme engineering, the rational modification of modular enzymes has been an important goal since their discovery. Early attempts to modify PKS and NRPS often resulted in significantly reduced or even inactive enzymes due to the exchange or deletion of some domains and modules. Evolutionary analysis of assembly line systems can infer sites where natural recombination occurs, guiding the design of artificial enzyme modification. Currently, new module definitions for PKS and the XUC concept for NRPS, derived from evolutionary analysis, provide a theoretical basis for the modification of modular enzymes. The application and further optimization of these concepts will drive the development of synthetic biology. Furthermore, this approach is not limited to multi-domain and multi-module enzymes like PKS and NRPS. Combining evolution with big data analysis will also provide new ideas for the modification of other enzymes.

Nature has created a rich and diverse array of biosynthetic pathways and natural products based on various evolutionary mechanisms over millions of years. Humanity's quest to understand nature has never ceased. By studying the evolutionary mechanisms of biosynthetic enzymes through bioinformatics, mining their active products for use in medicine, health, or agricultural production, combining big data analysis to depict a panoramic landscape of natural products, or utilizing nature's rules and components to design and modify biosynthetic enzymes from an evolutionary perspective to meet human needs, the process of discovering and transforming nature are engaged in.

Funding

National Natural Science Foundation of China (Project No.: 22177092, 82003631, 82104026); Zhejiang Provincial Leading Innovative and Entrepreneurial Team (Project No.: 2020R01004); Hangzhou Science and Technology Development Plan (Project No.: 20201203B122)

Disclosure statement

The authors declare no conflict of interest.

References

- [1] Chevrette MG, Gavrilidou A, Mantri S, et al., 2021, The Confluence of Big Data and Evolutionary Genome Mining for the Discovery of Natural Products. *Natural Product Reports*, 38(11): 2024–2040.
- [2] Chevrette MG, Gutierrez-Garcia K, Selem-Mojica N, et al., 2020, Evolutionary Dynamics of Natural Product Biosynthesis in Bacteria. *Natural Product Reports*, 37(4): 566–599.
- [3] Jensen PR, 2016, Natural Products and the Gene Cluster Revolution. *Trends in Microbiology*, 24(12): 968–977.
- [4] Gavrilidou A, Kautsar SA, Zaburanyi N, et al., 2022, Compendium of Specialized Metabolite Biosynthetic Diversity Encoded in Bacterial Genomes. *Nature Microbiology*, 7: 726–735.
- [5] Chen SC, Zhang C, Zhang LH, 2022, Investigation of the Molecular Landscape of Bacterial Aromatic Polyketides by Global Analysis of Type II Polyketide Synthases. *Angewandte Chemie-International Edition*, 61(24): e202202286.
- [6] Adamek M, Alanjary M, Ziemert N, 2019, Applied Evolution: Phylogeny-Based Approaches in Natural Products Research. *Natural Product Reports*, 36(9): 1295–1312.
- [7] Pande S, Kost C, 2017, Bacterial Unculturability and the Formation of Intercellular Metabolic Networks. *Trends in Microbiology*, 25(5): 349–361.
- [8] Ziemert N, Alanjary M, Weber T, 2016, The Evolution of Genome Mining in Microbes - A Review. *Natural Product Reports*, 33(8): 988–1005.
- [9] Walker J M, 2022, Engineering Natural Products Biosynthesis. Humana Imprint, New York.
- [10] Yang Q, Cheng BT, Tang ZJ, et al., 2021, Applications and Prospects of Genome Mining in the Discovery of Natural Products. *Synthetic Biology Journal*, 2(5): 697–715.
- [11] Zerikly M, Challis GL, 2009, Strategies for the Discovery of New Natural Products by Genome Mining. *ChemBioChem*, 10(4): 625–633.
- [12] Scherlach K, Hertweck C, 2021, Mining and Unearthing Hidden Biosynthetic Potential. *Nature Communications*, 12(1): 3864.
- [13] Bauman KD, Butler KS, Moore BS, et al., 2021, Genome Mining Methods to Discover Bioactive Natural Products. *Natural Product Reports*, 38(11): 2100–2129.
- [14] Cruz-Morales P, Kopp JF, Martinez-Guerrero C, et al., 2016, Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomyces. *Genome Biology and Evolution*, 8(6): 1906–1916.
- [15] Kang HS, 2017, Phylogeny-Guided (Meta)Genome Mining Approach for the Targeted Discovery of New Microbial Natural Products. *Journal of Industrial Microbiology and Biotechnology*, 44(2): 285–293.
- [16] Alanjary M, Kronmiller B, Adamek M, et al., 2017, The Antibiotic Resistant Target Seeker (ARTS), an Exploration Engine for Antibiotic Cluster Prioritization and Novel Drug Target Discovery. *Nucleic Acids Research*, 45(W1): W42–W48.
- [17] Mungan MD, Alanjary M, Blin K, et al., 2020, ARTS 2.0: Feature Updates and Expansion of the Antibiotic Resistant Target Seeker for Comparative Genome Mining. *Nucleic Acids Research*, 48(W1): W546–W552.
- [18] Ziemert N, Podell S, Penn K, et al., 2012, The Natural Product Domain Seeker NaPDos: A Phylogeny-Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. *PLoS One*, 7(3): e34064.
- [19] Klau LJ, Podell S, Creamer KE, et al., 2022, The Natural Product Domain Seeker Version 2 (NaPDos2) Webtool Relates Ketosynthase Phylogeny to Biosynthetic Function. *Journal of Biological Chemistry*, 298(10): 102480.
- [20] Selem-Mojica N, Aguilar C, Gutierrez-Garcia K, et al., 2019, EvoMining Reveals the Origin and Fate of Natural Product Biosynthetic Enzymes. *Microbial Genomics*, 5(12): e260.
- [21] Navarro-Munoz JC, Selem-Mojica N, Mullowney MW, et al., 2020, A Computational Framework to Explore Large-Scale Biosynthetic Diversity. *Nature Chemical Biology*, 16(1): 60–68.

- [22] Medema MH, Kottmann R, Yilmaz P, et al., 2015, Minimum Information about a Biosynthetic Gene Cluster. *Nature Chemical Biology*, 11(9): 625–631.
- [23] Kautsar SA, Blin K, Shaw S, et al., 2020, MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function. *Nucleic Acids Research*, 48(D1): D454–D458.
- [24] Blin K, Shaw S, Kautsar SA, et al., 2021, The AntiSMASH Database Version 3: Increased Taxonomic Coverage and New Query Features for Modular Enzymes. *Nucleic Acids Research*, 49(D1): D639–D643.
- [25] Rawlings BJ, 2001, Type I Polyketide Biosynthesis in Bacteria (Part A-Erythromycin Biosynthesis). *Natural Product Reports*, 18(2): 190–227.
- [26] Hertweck C, Luzhetskyy A, Rebets Y, et al., 2007, Type II Polyketide Synthases: Gaining a Deeper Insight into Enzymatic Teamwork. *Natural Product Reports*, 24(1): 162–190.
- [27] Abe I, Morita H, 2010, Structure and Function of the Chalcone Synthase Superfamily of Plant Type III Polyketide Synthases. *Natural Product Reports*, 27(6): 809–838.
- [28] Yu D, Xu F, Zeng J, et al., 2012, Type III Polyketide Synthases in Natural Product Biosynthesis. *IUBMB Life*, 64(4): 285–295.
- [29] Nivina A, Yuet KP, Hsu J, et al., 2019, Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chemical Reviews*, 119(24): 12524–12547.
- [30] Jenke-Kodama H, Dittmann E, 2009, Evolution of Metabolic Diversity: Insights from Microbial Polyketide Synthases. *Phytochemistry*, 70(15–16): 1858–1866.
- [31] Nivina A, Herrera PS, Fraser HB, et al., 2021, GRINS: Genetic Elements That Recode Assembly-Line Polyketide Synthases and Accelerate Their Diversification. *Proceedings of the National Academy of Sciences of the United States of America*, 118(26): e260.
- [32] Jenke-Kodama H, Sandmann A, Muller R, et al., 2005, Evolutionary Implications of Bacterial Polyketide Synthases. *Molecular Biology and Evolution*, 22(10): 2027–2039.
- [33] Nguyen T, Ishida K, Jenke-Kodama H, et al., 2008, Exploiting the Mosaic Structure of Trans-Acyltransferase Polyketide Synthases for Natural Product Discovery and Pathway Dissection. *Nature Biotechnology*, 26(2): 225–233.
- [34] Lopez JV, 2004, Naturally Mosaic Operons for Secondary Metabolite Biosynthesis: Variability and Putative Horizontal Transfer of Discrete Catalytic Domains of the Epothilone Polyketide Synthase Locus. *Molecular Genetics and Genomics*, 270(5): 420–431.
- [35] Zhang L, Hashimoto T, Qin B, et al., 2017, Characterization of Giant Modular PKSs Provides Insight into Genetic Mechanism for Structural Diversification of Aminopolyol Polyketides. *Angewandte Chemie - International Edition*, 56(7): 1740–1745.
- [36] Keatinge-Clay AT, 2017, Polyketide Synthase Modules Redefined. *Angewandte Chemie - International Edition*, 56(17): 4658–4660.
- [37] Vander WDA, Keatinge-Clay AT, 2018, The Modules of Trans-Acyltransferase Assembly Lines Redefined with a Central Acyl Carrier Protein. *Proteins*, 86(6): 664–675.
- [38] Caffrey P, 2003, Conserved Amino Acid Residues Correlating with Ketoreductase Stereospecificity in Modular Polyketide Synthases. *Chembiochem*, 4(7): 654–657.
- [39] Drew AVW, Adrian TKC, 2018, The Modules of Trans-Acyltransferase Assembly Lines Redefined with a Central Acyl Carrier Protein. *Proteins*, 86(6): 664–675.
- [40] Medema MH, Cimermanic P, Sali A, et al., 2014, A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. *PLoS Computational Biology*, 10(12): e1004016.
- [41] Ridley CP, Lee HY, Khosla C, 2008, Evolution of Polyketide Synthases in Bacteria. *Proceedings of the National Academy*

- of Sciences of the United States of America, 15(12): 4595–4600.
- [42] Hillenmeyer ME, Vandova GA, Berlew EE, et al., 2015, Evolution of Chemical Diversity by Coordinated Gene Swaps in Type II Polyketide Gene Clusters. *Proceedings of the National Academy of Sciences of the United States of America*, 112(45): 13952–13957.
- [43] Gabaldon T, Koonin EV, 2013, Functional and Evolutionary Implications of Gene Orthology. *Nature Reviews Genetics*, 14(5): 360–366.
- [44] Fritzsche K, Ishida K, Hertweck C, 2008, Orchestration of Discoid Polyketide Cyclization in the Resistomycin Pathway. *Journal of the American Chemical Society*, 130: 8307–8316.
- [45] Fraley AE, Dieterich CL, Mabesoone MFJ, et al., 2022, Structure of a Promiscuous Thioesterase Domain Responsible for Branching Acylation in Polyketide Biosynthesis. *Angewandte Chemie - International Edition*, 2022: e202206385.
- [46] Schwecke T, Aparicio JF, Molnár I, et al., 1995, The Biosynthetic Gene Cluster for the Polyketide Immunosuppressant Rapamycin. *Proceedings of the National Academy of Sciences of the United States of America*, 92(17): 7839–7843.
- [47] Haydock SF, Aparicio JF, Molnár I, et al., 1995, Divergent Sequence Motifs Correlated with the Substrate Specificity of (Methyl) Malonyl-CoA: Acyl Carrier Protein Transacylase Domains in Modular Polyketide Synthases. *FEBS Letters*, 374(2): 246–248.
- [48] Aparicio JF, Molnár I, Schwecke T, et al., 1996, Organization of the Biosynthetic Gene Cluster for Rapamycin in *Streptomyces hygroscopicus*: Analysis of the Enzymatic Domains in the Modular Polyketide Synthase. *Gene*, 169(1): 9–16.
- [49] Kakavas SJ, Katz L, Stassi D, 1997, Identification and Characterization of the Niddamycin Polyketide Synthase Genes from *Streptomyces caelestis*. *Journal of Bacteriology*, 179(23): 7515–7522.
- [50] Kang HS, Brady SF, 2014, Mining Soil Metagenomes to Better Understand the Evolution of Natural Product Structural Diversity: Pentangular Polyphenols as a Case Study. *Journal of the American Chemical Society*, 136(52): 18111–18119.
- [51] Kang HS, Brady SF, 2013, Arimetamycin A: Improving Clinically Relevant Families of Natural Products Through Sequence-Guided Screening of Soil Metagenomes. *Angewandte Chemie - International Edition*, 52(42): 11063–11067.
- [52] Li LY, Hu YL, Sun JL, et al., 2022, Resistance and Phylogeny-Guided Discovery Reveals Structural Novelty of Tetracycline Antibiotics. *Chemical Science*, 13: 12892–12898.
- [53] Ziemert N, Jensen PR, 2012, Phylogenetic Approaches to Natural Product Structure Prediction. *Methods in Enzymology*, 517: 161–182.
- [54] Ueoka R, Uria AR, Reiter S, et al., 2015, Metabolic and Evolutionary Origin of Actin-Binding Polyketides from Diverse Organisms. *Nature Chemical Biology*, 11(9): 705–712.
- [55] Helfrich EJN, Ueoka R, Dolev A, et al., 2019, Automated Structure Prediction of Trans-Acyltransferase Polyketide Synthase Products. *Nature Chemical Biology*, 15(8): 813–821.
- [56] Helfrich EJN, Ueoka R, Chevrette MG, et al., 2021, Evolution of Combinatorial Diversity in Trans-Acyltransferase Polyketide Synthase Assembly Lines Across Bacteria. *Nature Communications*, 12(1): 1422.
- [57] Guo J, Ran HM, Zeng J, et al., 2016, Tafuketide, a Phylogeny-Guided Discovery of a New Polyketide from *Talaromyces funiculosus* Salicorn 58. *Applied Microbiology and Biotechnology*, 100: 5323–5338.
- [58] Shen B, Hindra, Yan X, et al., 2015, Eneidyne: Exploration of Microbial Genomics to Discover New Anticancer Drug Leads. *Bioorganic and Medicinal Chemistry Letters*, 25(1): 9–15.
- [59] Yan X, Ge H, Huang T, et al., 2016, Strain Prioritization and Genome Mining for Eneidyne Natural Products. *mBio*, 7(6): 12.
- [60] Weissman KJ, 2016, Genetic Engineering of Modular PKSs: From Combinatorial Biosynthesis to Synthetic Biology. *Natural Product Reports*, 33(2): 203–230.
- [61] Cao CK, Li JL, Zhang KC, 2021, Research Progress in Synthetic Organic Alcohols and Organic Acids Through Artificial

- Metabolic Pathways. *Synthetic Biology*, 2(6): 902–919.
- [62] Booth TJ, Bozhuyuk KAJ, Liston JD, et al., 2022, Bifurcation Drives the Evolution of Assembly-Line Biosynthesis. *Nature Communications*, 13(1): 3498.
- [63] Hirsch M, Fitzgerald BJ, Keatinge-Clay AT, 2021, How Cis-Acyltransferase Assembly-Line Ketosynthases Gatekeep for Processed Polyketide Intermediates. *ACS Chemical Biology*, 16(11): 2515–2526.
- [64] Miyazawa T, Hirsch M, Zhang ZC, et al., 2020, An *In Vitro* Platform for Engineering and Harnessing Modular Polyketide Synthases. *Nature Communications*, 11(1): 80.
- [65] Miyazawa T, Fitzgerald BJ, Keatinge-Clay AT, 2021, Preparative Production of an Enantiomeric Pair by Engineered Polyketide Synthases. *Chemical Communications*, 57(70): 8762–8765.
- [66] Peng H, Ishida K, Sugimoto Y, et al., 2019, Emulating Evolutionary Processes to Morph Aureothin-Type Modular Polyketide Synthases and Associated Oxygenases. *Nature Communications*, 10(1): 3918.
- [67] Yuzawa S, Deng K, Wang G, et al., 2017, Comprehensive In Vitro Analysis of Acyltransferase Domain Exchanges in Modular Polyketide Synthases and Its Application for Short-Chain Ketone Production. *ACS Synthetic Biology*, 6(1): 139–147.
- [68] Satoshi Y, Mona M, Renee J, et al., 2018, Short-Chain Ketone Production by Engineered Polyketide Synthases in *Streptomyces albus*. *Nature Communications*, 9(1): 4569.
- [69] Zargar A, Valencia L, Wang J, et al., 2020, A Bimodular PKS Platform That Expands the Biological Design Space. *Metabolic Engineering*, 61: 389–396.
- [70] Eng CH, Yuzawa S, Wang G, et al., 2016, Alteration of Polyketide Stereochemistry From Anti to Syn by a Ketoreductase Domain Exchange in a Type I Modular Polyketide Synthase Subunit. *Biochemistry*, 55(12): 1677–1680.
- [71] Hagen A, Poust S, De Rond T, et al., 2016, Engineering a Polyketide Synthase for In Vitro Production of Adipic Acid. *ACS Synthetic Biology*, 5(1): 21–27.
- [72] Zargar A, Lal R, Valencia L, et al., 2020, Chemoinformatic-Guided Engineering of Polyketide Synthases. *Journal of the American Chemical Society*, 142(22): 9896–9901.
- [73] Musiol-Kroll EM, Wohlleben W, 2018, Acyltransferases as Tools for Polyketide Synthase Engineering. *Antibiotics-Basel*, 7(3): 62.
- [74] Kwan DH, Tosin M, Schlager N, et al., 2011, Insights Into the Stereospecificity of Ketoreduction in a Modular Polyketide Synthase. *Organic & Biomolecular Chemistry*, 9(7): 2053–2056.
- [75] Bagde SR, Mathews II, Fromme JC, et al., 2021, Modular Polyketide Synthase Contains Two Reaction Chambers That Operate Asynchronously. *Science*, 374(6568): 723.
- [76] Cogan DP, Zhang KM, Li XY, et al., 2021, Mapping the Catalytic Conformations of an Assembly-Line Polyketide Synthase Module. *Science*, 374(6568): 729–734.
- [77] Herbst DA, Jakob RP, Zähringer F, et al., 2016, Mycocerosic Acid Synthase Exemplifies the Architecture of Reducing Polyketide Synthases. *Nature*, 531(7595): 533–537.
- [78] Zheng JT, Gay DC, Demeler B, et al., 2012, Divergence of Multimodular Polyketide Synthases Revealed by a Didomain Structure. *Nature Chemical Biology*, 8(7): 615–621.
- [79] Gay D, You YO, Keatinge-Clay A, et al., 2013, Structure and Stereospecificity of the Dehydratase Domain from the Terminal Module of the Rifamycin Polyketide Synthase. *Biochemistry*, 52(49): 8916–8928.
- [80] Dutta S, Whicher JR, Hansen DA, et al., 2014, Structure of a Modular Polyketide Synthase. *Nature*, 510(7506): 512–517.
- [81] Whicher JR, Dutta S, Hansen DA, et al., 2014, Structural Rearrangements of a Polyketide Synthase Module During Its Catalytic Cycle. *Nature*, 510(7506): 560–564.
- [82] Felnagle EA, Jackson EE, Chan YA, et al., 2008, Nonribosomal Peptide Synthetases Involved in the Production of

- Medically Relevant Natural Products. *Molecular Pharmacology*, 5(2): 191–211.
- [83] Sieber SA, Marahiel MA, 2005, Molecular Mechanisms Underlying Nonribosomal Peptide Synthesis: Approaches to New Antibiotics. *Chemical Reviews*, 105(2): 715–738.
- [84] Süssmuth RD, Mainz A, 2017, Nonribosomal Peptide Synthesis: Principles and Prospects. *Angewandte Chemie-International Edition*, 56(14): 3770–3821.
- [85] Jaremko MJ, Davis TD, Corpuz JC, et al., 2020, Type II Nonribosomal Peptide Synthetase Proteins: Structure, Mechanism, and Protein-Protein Interactions. *Natural Product Reports*, 37(3): 355–379.
- [86] Calcott MJ, Owen JG, Ackerley DF, 2020, Efficient Rational Modification of Non-Ribosomal Peptides by Adenylation Domain Substitution. *Nature Communications*, 11(1): 4554.
- [87] Baunach M, Chowdhury S, Stallforth P, et al., 2021, The Landscape of Recombination Events That Create Nonribosomal Peptide Diversity. *Molecular Biology and Evolution*, 38(5): 2116–2130.
- [88] Rausch C, Hoof I, Weber T, et al., 2007, Phylogenetic Analysis of Condensation Domains in NRPS Sheds Light on Their Functional Evolution. *BMC Ecology and Evolution*, 7: 78.
- [89] Wheadon MJ, Townsend CA, 2021, Evolutionary and Functional Analysis of an NRPS Condensation Domain Integrates Beta-Lactam, -Amino Acid, and Dehydroamino Acid Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 118(17): e2026017118.
- [90] Stachelhaus T, Mohamed A, Marahiel MA, 1999, The Specificity-Confering Code of Adenylation Domains in Nonribosomal Peptide Synthetases. *Chemistry & Biology*, 6: 493–505.
- [91] Röttig M, Medema MH, Blin K, et al., 2011, NRPSpredictor2 - A Web Server for Predicting NRPS Adenylation Domain Specificity. *Nucleic Acids Research*, 39: W362–W367.
- [92] Chevrette MG, Aicheler F, Kohlbacher O, et al., 2017, SANDPUMA: Ensemble Predictions of Nonribosomal Peptide Chemistry Reveal Biosynthetic Diversity Across Actinobacteria. *Bioinformatics*, 33(20): 3202–3210.
- [93] Ziemert N, Lechner A, Wietz M, et al., 2014, Diversity and Evolution of Secondary Metabolism in the Marine Actinomycete Genus *Salinispora*. *Proceedings of the National Academy of Sciences of the United States of America*, 111(12): 1130–1139.
- [94] Patteson J B, Fortinez C M, Putz A T, et al., 2022, Structure and Function of a Dehydrating Condensation Domain in Nonribosomal Peptide Biosynthesis. *Journal of the American Chemical Society*, 144(31): 14057–14070.
- [95] Hover BM, Kim SH, Katz M, et al., 2018, Culture-Independent Discovery of the Malacidins as Calcium-Dependent Antibiotics With Activity Against Multidrug-Resistant Gram-Positive Pathogens. *Nature Microbiology*, 3(4): 415–422.
- [96] Culp EJ, Waglechner N, Wang W, et al., 2020, Evolution-Guided Discovery of Antibiotics That Inhibit Peptidoglycan Remodeling. *Nature*, 578(7796): 582–587.
- [97] Xu M, Wang WL, Waglechner N, et al., 2022, Phylogeny-Informed Synthetic Biology Reveals Unprecedented Structural Novelty in Type V Glycopeptide Antibiotics. *ACS Central Science*, 8(5): 615–626.
- [98] Wang ZQ, Koirala B, Hernandez Y, et al., 2022, Bioinformatic Prospecting and Synthesis of a Bifunctional Lipopeptide Antibiotic That Evades Resistance. *Science*, 376: 991–996.
- [99] Calcott MJ, Owen JG, Lamont IL, et al., 2014, Biosynthesis of Novel Pyoverdines by Domain Substitution in a Nonribosomal Peptide Synthetase of *Pseudomonas aeruginosa*. *Applied and Environmental Microbiology*, 80(18): 5723–5731.
- [100] Thirlway J, Lewis R, Nunns L, et al., 2012, Introduction of a Non-Natural Amino Acid Into a Nonribosomal Peptide Antibiotic by Modification of Adenylation Domain Specificity. *Angewandte Chemie-International Edition*, 51(29): 7181–7184.
- [101] Kries H, Wachtel R, Pabst A, et al., 2014, Reprogramming Nonribosomal Peptide Synthetases for “Clickable” Amino

- Acids. *Angewandte Chemie-International Edition*, 53(38): 10105–10108.
- [102] Kien T, Nguyen DR, Gu JQ, et al., 2006, Combinatorial Biosynthesis of Novel Antibiotics Related to Daptomycin. *Proceedings of the National Academy of Sciences of the United States of America*, 103(46): 17462–17467.
- [103] Crusemann M, Kohlhaas C, Piel J., 2013, Evolution-Guided Engineering of Nonribosomal Peptide Synthetase Adenylation Domains. *Chemical Science*, 4(3): 1041–1045.
- [104] Kries H, Niquille DL, Hilvert D., 2015, A Subdomain Swap Strategy for Reengineering Nonribosomal Peptides. *Chemistry & Biology*, 22(5): 640–648.
- [105] Bozhüyük KAJ, Linck A, Tietze A, et al., 2019, Modification and De Novo Design of Non-Ribosomal Peptide Synthetases (NRPS) Using Specific Assembly Points Within Condensation Domains. *Nature Chemistry*, 11: 653–661.
- [106] Bozhüyük KAJ, Watzel J, Abbood N, et al., 2021, Synthetic Zippers as an Enabling Tool for Engineering of Non-Ribosomal Peptide Synthetases. *Angewandte Chemie-International Edition*, 60(32): 17531–17538.
- [107] Kranz J, Wenski SL, Dichter AA, et al., 2021, Influence of Condensation Domains on Activity and Specificity of Adenylation Domains. *bioRxiv*, 2021: 1–45.
- [108] Bozhüyük K A J, Fleischhacker F, Linck A, et al., 2018, De Novo Design and Engineering of Non-Ribosomal Peptide Synthetases. *Nature Chemistry*, 10(3): 275–281.
- [109] Minami A, Ugai T, Ozaki T, et al., 2020, Predicting the Chemical Space of Fungal Polyketides by Phylogeny-Based Bioinformatics Analysis of Polyketide Synthase-Nonribosomal Peptide Synthetase and Its Modification Enzymes. *Scientific Reports*, 10: 13556.
- [110] Awakawa T, Fujioka T, Zhang L, et al., 2018, Reprogramming of the Antimycin NRPS-PKS Assembly Lines Inspired by Gene Evolution. *Nature Communications*, 9(1): 3534.
- [111] Santos-Aberturas J, Chandra G, Frattaruolo L, et al., 2019, Uncovering the Unexplored Diversity of Thioamidated Ribosomal Peptides in Actinobacteria Using the RiPPER Genome Mining Tool. *Nucleic Acids Research*, 47(9): 4624–4637.
- [112] Lü JW, Deng ZX, Zhang Q, et al., 2022, Identification of RiPPs Precursor Peptides and Cleavage Sites Based on Deep Learning. *Synthetic Biology*, 2022: 1–14.
- [113] Medema MH, Takano E, Breitling R., 2013, Detecting Sequence Homology at the Gene Cluster Level With MultiGeneBlast. *Molecular Biology and Evolution*, 30: 1218–1223.
- [114] Tietz JI, Schwalen CJ, Patel PS, et al., 2017, A New Genome-Mining Tool Redefines the Lasso Peptide Biosynthetic Landscape. *Nature Chemical Biology*, 13(5): 470–475.
- [115] Merwin NJ, Mousa WK, Dejong CA, et al., 2020, DeepRiPP Integrates Multiomics Data to Automate Discovery of Novel Ribosomally Synthesized Natural Products. *Proceedings of the National Academy of Sciences of the United States of America*, 117(1): 371–380.
- [116] Martin-Sanchez L, Singh KS, Avalos M, et al., 2019, Phylogenomic Analyses and Distribution of Terpene Synthases Among *Streptomyces*. *Beilstein Journal of Organic Chemistry*, 15: 1181–1193.
- [117] Jia Q, Chen X, Köllner TG, et al., 2019, Terpene Synthase Genes Originated From Bacteria Through Horizontal Gene Transfer Contribute to Terpenoid Diversity in Fungi. *Scientific Reports*, 9(1): 9223.
- [118] Avalos M, Garbeva P, Vader L, et al., 2022, Biosynthesis, Evolution, and Ecology of Microbial Terpenoids. *Natural Product Reports*, 39(2): 249–272.
- [119] Yang YL, Zhang SS, Ma K, et al., 2017, Discovery and Characterization of a New Family of Diterpene Cyclases in Bacteria and Fungi. *Angewandte Chemie-International Edition*, 56(17): 4749–4752.
- [120] Chen R, Jia QD, Mu X, et al., 2021, Systematic Mining of Fungal Chimeric Terpene Synthases Using an Efficient Precursor-Providing Yeast Chassis. *Proceedings of the National Academy of Sciences of the United States of America*,

118(29): e2023247118.

- [121] Tao H, Lauterbach L, Bian GK, et al., 2022, Discovery of Non-Squalene Triterpenes. *Nature*, 606: 414–420.
- [122] Jiang C, Kim SY, Suh DY., 2008, Divergent Evolution of the Thiolase Superfamily and Chalcone Synthase Family. *Molecular Phylogenetics and Evolution*, 49(3): 691–701.
- [123] Tan Z, Clomburg JM, Cheong S, et al., 2020, A Polyketoacyl-CoA Thiolase-Dependent Pathway for the Synthesis of Polyketide Backbones. *Nature Catalysis*, 3(7): 593–603.
- [124] Shanklin J, Guy JE, Mishra G, et al., 2009, Desaturases: Emerging Models for Understanding Functional Diversification of Diiron-Containing Enzymes. *Journal of Biological Chemistry*, 284(28): 18559–18563.
- [125] Zhu X, Liu J, Zhang W, 2015, De novo Biosynthesis of Terminal Alkyne-Labeled Natural Products. *Nature Chemical Biology*, 11(2): 115–120.
- [126] Zhu X, Su M, Manickam K, et al., 2015, Bacterial Genome Mining of Enzymatic Tools for Alkyne Biosynthesis. *ACS Chemical Biology*, 10(12): 2785–2793.
- [127] Chang FY, Brady SF, 2013, Discovery of Indolotryptoline Antiproliferative Agents by Homology-Guided Metagenomic Screening. *Proceedings of the National Academy of Sciences of the United States of America*, 110(7): 2478–2483.
- [128] Chang FY, Ternei MA, Calle PY, et al., 2015, Targeted Metagenomics: Finding Rare Tryptophan Dimer Natural Products in the Environment. *Journal of the American Chemical Society*, 137(18): 6044–6052.
- [129] Cimermancic P, Medema MH, Claesen J, et al., 2014, Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell*, 158(2): 412–421.
- [130] O'Neill EC, Schorn M, Larson CB, et al., 2019, Targeted Antibiotic Discovery through Biosynthesis-Associated Resistance Determinants: Target-Directed Genome Mining. *Critical Reviews in Microbiology*, 45(3): 255–277.
- [131] Yan Y, Liu N, Tang Y, 2020, Recent Developments in Self-Resistance Gene-Directed Natural Product Discovery. *Natural Product Reports*, 37(7): 879–892.
- [132] Bernhardsgrutter I, Schell K, Peter DM, et al., 2019, Awakening the Sleeping Carboxylase Function of Enzymes: Engineering the Natural CO₂-Binding Potential of Reductases. *Journal of the American Chemical Society*, 141(25): 9778–9782.
- [133] Sikosek T, 2019, Computational Methods in Protein Evolution. *Methods in Molecular Biology*, Humana Press, 2019: 1064–3745.
- [134] Harms MJ, Thornton JW, 2010, Analyzing Protein Structure and Function Using Ancestral Gene Reconstruction. *Current Opinion in Structural Biology*, 20(3): 360–366.
- [135] Cech NB, Medema MH, Clardy J, 2021, Benefiting from Big Data in Natural Products: Importance of Preserving Foundational Skills and Prioritizing Data Quality. *Natural Product Reports*, 38(11): 1947–1953.
- [136] Jeon J, Kang S, Kim HU, 2021, Predicting Biochemical and Physiological Effects of Natural Products from Molecular Structures Using Machine Learning. *Natural Product Reports*, 38(11): 1954–1966.

Publisher's note

Whioce Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.