**Discovery of novel bioactive natural products driven by genome mining**

Zhongyue Li¹, Deyu Zhu², Yuemao Shen¹,*

¹ Key Laboratory of Chemical Biology, School of Pharmaceutical Sciences, Shandong University, Ji’nan, Shandong, China; ² School of Basic Medical Sciences, Shandong University, Ji’nan, Shandong, China.

### Summary

Microbial natural product is an important source for drug discovery. As more and more microbial genomes are sequenced, bioinformatics analysis shows that there are huge resources of novel natural products. Genome mining is a new strategy of natural product discovery based on gene cluster sequences and biosynthetic pathways. At the same time, it can directly associate the structures of natural products with synthetic pathways, and facilitate the study of biosynthesis and combinatorial biosynthesis. In this paper, the strategies of genome mining, including bioinformatics predictions, metabolomic comparisons and genetic manipulations, are reviewed, which shows a great advantage of this strategy in exploiting the potential of microbial natural products. With the development of genome mining methodology and techniques, it will be possible to realize rational exploitation of microbial natural product resources.

**Keywords:** Genome mining, natural products, gene cluster, biosynthesis; microorganisms

---

### 1. Introduction

The discovery of streptomycin from *Streptomyces* has inspired the research of antibiotics from microorganisms. The discovery rate of new antibiotics has been increasing year by year, reaching its peak in the 1970s, and experienced a golden period of nearly 30 years. However, in the 1980s, due to the repeated isolation and screening of most actinomycetes, and the repeated isolation of bioactive natural products, the probability of finding new active compounds decreased rapidly and became more and more difficult. However, according to the most conservative estimates, the number of natural products found so far from any microbial strain is only 20% of the total encoded by its genome. Therefore, the decline in the rate of discovery of new active natural products is not due to the exhaustion of the biosynthetic gene resources of microbial natural products, but is limited by the methods of microbial isolation, cultivation and bioactivity screening (1). More importantly, since the birth of microbial natural product chemistry, the research methodology adapted has striking randomness, blindness and contingency, is insufficient in rationality. There is a lack of rational experimental design and theoretical guidance from isolation and culture of strains to extraction and separation of active components in addition to activity screening, so we can only expect to obtain the activity specified in the screening model but can't predict the structures of bioactive components.

In recent years, with decreasing sequencing costs and continuous improvements in sequencing and longer read technologies, more and more microorganism genomes have been sequenced. The emergence of these genomic information is gradually changing the research model of the whole life sciences and related disciplines. It is found that the natural product resources of conventional microorganisms are greatly underestimated (2). Taking bacterial genomes as an example, bioinformatics analysis shows that each strain of bacteria has the ability to synthesize 25-30 types of natural products, of which ca. 90% is still unknown. Even if only 20% of the 25-30 types of natural products produced by each strain are new structures, the number of new natural products we obtain will increase exponentially (3).
The Genomes Online Database (GOLD v.6) (https://gold.jgi.doe.gov) shows huge genomic data (4), which imposes a challenging and exciting research topic how to obtain novel natural products quickly and effectively by the method of genome mining. Compared with traditional methods, novel natural product discovery strategies based on gene cluster sequences and biosynthesis pathways have the following advantages: 1. Through gene "prediction" of natural product structure, "de-duplication" can be achieved. The possibility of finding new structures would be greatly increased by excluding known compounds. 2. Combined with transcriptional analysis and genetic manipulation of regulatory genes, the secondary metabolic potential of the strain could be released to the maximal extent. 3. The structures of natural products can be directly linked with the synthetic pathways, which can facilitate the study of biosynthesis and combinatorial biosynthesis, and further expand the structural diversity of natural products. Thus far, natural product discovery has been transformed to heavily rely on bioinformatics-based structure prediction, activation of cryptic gene clusters by genetic manipulation and heterologous expression, and potential engineering of natural product analogs (5).

2. Discovery of novel natural products with the aid of bioinformatics prediction and comparative metabolomic analysis

Until the late 1980s, biosynthetic pathways of microbial natural products could only be inferred by feeding selectively labelled precursors, and by analyzing the MS and NMR data of the products to obtain the labelling patterns of the products. In 1984, Malpartida and Hopwood reported the cloning and heterologous expression of the biosynthetic gene cluster of actinorhodin (6). Katz and his colleagues from Abbott and Leadlay from Cambridge University reported independently the cloning of polyketide synthase (PKS) gene of 6-deoxyerythronolide B (6-DEB) in 1990 and 1991, respectively, which greatly changed the research mode of biosynthetic pathway of natural products. Subsequently, it was found that the nonribosomal peptides such as penicillin/cephalosporin, vancomycin, cyclosporin and daptomycin were also synthesized by the assembly lines, nonribosomal peptide synthetases (NRPSs), similar to that of 6-DEB PKS. These studies have changed our understanding of biosynthetic mechanisms of peptide natural products. With the further study of biosynthetic genes of rapamycin and FK506, it has been found that their syntheses have the compatibility of biosynthetic system with assembly line characteristic by hybrid of PKS and NRPS. This seemingly simple "assembly line enzymology" can explain the syntheses of thousands of complex and diverse natural products.

It is precisely because of the discovery of those biosynthetic pathways, the number of modules contained and the composition of the domain of each module can be determined by analyzing the sequences of PKS and NRPS genes of multiple modules, and the compound structures encoded by them can be predicted. By analyzing the predicted structures of the compounds, we can identify the assembly modules and the related physicochemical properties needed for the synthesis, so as to direct the isolation procedure and finally obtain the natural products encoded by the corresponding biosynthetic gene clusters (7). Specifically, the genomisotopic strategy includes discovering novel natural products on the basis of predicting structural moieties and biosynthetic precursors (8), which is carried out by bioinformatics analysis, subsequently, feeding isotope labelled precursors and isotope-guided isolation of the targeted natural products. E.g., orfamides A-C were isolated from *Pseudomonas fluorescens* Pf-5 by this genomisotopic approach as bioinformatics analysis indicated that these cyclopeptides contain four L-leucine residues (Figure 1) (8). Erythrochelin was obtained from *Saccharopolyspora erythraea* by feeding 14C-L-ornithine followed by radio-LC-MS-guided isolation (Figure 1) (9,10). Additionally, bioinformatics analysis revealed that the cch gene cluster could be responsible...
for the biosynthesis of the tetrapeptides with the presence of a hydroxamate moiety. Hydroxamate readily conjugates with ferric ion, which produces characteristic UV absorbance. Through UV comparison of the products of Streptomyces coelicolor M145 mutant and wild type strains, coelichelin was obtained (11). UV spectra-guided isolations were also successful in the discovery of many chromophore-containing natural products, e.g., aspoquinolones A-D from Aspergillus nidulans (12), salinilactam A from Salinispora tropica (13), thailandamide A from Burkholderia thailandensis (14). The feasibility of structural moiety-based genome mining was demonstrated in screening a collection of over 10,000 actinomycetes for the genetic potential of producing phosphonic acids. Before this study no genome mining experiments were performed on a large number of bacterial strains (15).

For the natural products encoded by unknown gene clusters in genomes, it is difficult to predict their structures or physiochemical properties (16). Usually, it is useful to obtain these natural products through the approach of gene inactivation and metabolomic comparison of the mutant and the wild type strains by HPLC or LC-MS analysis (17-20). HPLC profile comparison is the most straightforward method used in genome mining. However, its limitation is obvious in that low resolution and rough identities of products. Once coupled with bioassays, this method could be feasible. WAP-8294As are cyclic depsipeptides and potent anti-MRSA agents. Their biosynthetic genes were not reported before Zhang et al.’s study (21). The genome of Lysobacter enzymogenes OH11 contains a large number of biosynthetic gene clusters including several hybrid PKS-NRPS ones. Disrupting one of the gene clusters containing two huge open reading frames, together encoding 12 modules of NRPS, resulted in the HPLC profile change and disappearance of the antibacterial activity of the mutant. Finally, Zhang et al. obtained WAP-8294A2 through HPLC- and bioassay-guided fractionation, so to characterize the biosynthetic gene cluster for this type of depsipeptides (21).

For the discovery of peptide natural products, an improved MS-guided genome mining method called NPP (natural product peptidogenomics) was invented by iteratively matching de novo tandem MS structures to genomics-based structures. Indeed, NPP enabled the rapid characterization of diverse ribosomal and nonribosomal peptide natural products (22). Recently, the MS-guided genome mining method has been evolved to molecular networking, a MS/MS data organizational approach (23), which can be integrated with many other methods, e.g., bioassays (24) and chemical synthesis (25). To simultaneously assess large numbers of closely related bacterial strains and their pan-metabolomes, molecular networking should be improved to recognize the MS/MS fragmentation patterns with the availability of biosynthetic gene clusters. This newly invented method identifies molecular families rather than individual molecules thus that enables the discovery of related molecules. Duncan et al. invented and applied this method to thirty five Salinispora strains, including 30 with draft genome sequences, which successfully resulted in the isolation and characterization of a quinomycin-type depsipeptide, namely retimycin A (26).

3. Activating cryptic gene clusters through genetic manipulations

In the genome of actinomycetes, many secondary metabolic gene clusters are silent. If we know how to activate these clusters, we will discover huge number of new enzymes, biological pathways and natural product resources (27, 28). Microbial secondary metabolism has a very complex transcriptional regulatory system (29), probably because: 1. The biosynthetic gene cluster is composed of multiple genes of different origins containing independent regulatory elements. 2. The combined biosynthetic pathway can use different primary metabolites as the starting substances for synthesis, which requires a complex regulatory system. 3. Having more than one metabolic regulation system can help the secondary metabolism process to be sensitive to the external environment changes. In particular, the responses of actinomycetes to environmental factors are more complex than other bacteria. Many regulatory factors are involved in the secondary metabolism regulation of actinomycetes and form a complex multi-level network system. These include high-level multifunctional genes and low-level pathway-specific genes.

Activating cryptic gene clusters can be carried out through manipulating global regulations including: 1. Optimization of culture conditions including the composition of culture media, aeration volume, shape of fermenter, light condition, culture temperature, adding enzyme inhibitors (30), elicitors (31), rare earth elements (32), and co-culture (33). 2. Ribosomal engineering, a breeding method of ribosomal mutant, which uses various resistant mutants of microorganisms as screening markers to obtain highly efficient ribosomal mutants with improved secondary metabolite biosynthesis (34,35). The methods of activating cryptic gene clusters through manipulating pathway specific regulations include: 1. Overexpression of pathway specific transcriptional factors. Structural genes such as PKS genes and NRPS genes in the secondary metabolic gene cluster of actinomycetes are usually concentrated on 1-2 operons. The wild type promoters can be replaced by gene recombination with constitutive strong promoters such as PermE, which bypasses the transcriptional regulations in wild-type strains and activate or partially activate the corresponding gene clusters. 2. Heterologous expression of putative cryptic...
gene clusters (14,36). The exploitation of secondary metabolic gene clusters requires the continuous development of transgenic cluster techniques among different organisms, including the optimization of super-hosts and the improvement of the methods for constructing/expressing multiple genes in heterologous systems (37,38).

4. Transcriptional regulations of secondary metabolisms in actinomycetes

The transcriptional regulations are particularly complex in actinomycetes. Multifunctional global regulators not only regulate the production of multiple secondary metabolites, but also participate in the morphological differentiation of actinomycetes, such as BldA, A-factors, Sigma factor and two-component regulatory systems often cascade into complex regulatory networks. Path-specific regulatory factors are mainly involved in the biosynthesis of specific metabolites, such as members of transcription factor families such as SARP (Streptomycetes antibiotic regulatory protein) and LuxR. These specific regulatory factors are often the only way for various global regulators to regulate specific biosynthetic pathways.

Among the numerous families of transcriptional regulatory proteins of actinomycetes (39), TetR (Tetracycline Repressor protein) family plays an important role in coordinating the growth and development of actinomycetes and responses to environmental factors. TetR family members are mainly involved in the regulation of antibiotic biosynthesis, efflux pumps and osmotic stress, and particularly abundant in environmentally vulnerable microorganisms, e.g. Streptomyces coelicolor A3(2) and S. avermitilis MA-4680 have 150 and 116 TetR members, respectively (40). TetRs are double domain proteins, one domain binds inducer and receives signals, another domain binds promoter DNA and transduces signals. Only the DNA binding domains are conserved among the TetRs, the elicitor binding domains are not significantly conserved, indicating the fine-tuned responses of actinomycetes to the high diversity of the inducers in the environments. After the unconserved domains binding inducers, the DNA binding domains may be altered allosterically and the ability of binding to DNA be lost, thus releasing the transcriptional repressing effects of TetRs. The TetR-binding inducers, also known as self-regulatory factors, usually are small molecular compounds that initiate secondary metabolism and morphological differentiation at the concentrations of nanomolar, and act as the hormones in eukaryotes.

Presently known self-regulatory factors of actinomycetes include 2,3-disubstituted γ-butyrolactone and its analogues including A-factor, SCB1, 2 and 3, VBA, B, C, D and E, IM-2 and Factor I, methylenomycin furans (41), P1 factor (42), peptides including SapB, goadsporin and Factor C, siderophores and certain antibiotics (43,44). A-factor was originally found in S. griseus (45,46). The receptor protein of A-factor is ArpA. When A-factor binds to ArpA, the transcriptional repression of ArpA is removed (47). The heterologous expression of ArpA always produces inclusion bodies. The CprB in S. coelicolor A3(2) is the homologous protein of ArpA, and the crystal structure is obtained. However, the small molecular ligands of CprB have not been found (48).

Because self-regulatory factors are similar to the "second messengers" of environmental signals (stimuli), it may be possible to integrate the regulation of secondary metabolisms and morphological differentiations of actinomycetes (integrative regulations). This mechanism is widely used in the regulation of secondary metabolism by actinomycetes, which is consistent with that secondary metabolisms are environment sensitive. Therefore, identification of elicitors and related regulatory mechanisms is another way to activate cryptic gene clusters in actinomycetes (49).

5. Transcriptional regulations of ansamycin biosynthesis

The ansamycin class of natural products includes the important anti-tuberculosis drug rifamycin (50), and the antitumor drug leads geldanamycins (51) and maytansinoids (52). These macrolactams arise from the polyketide pathway via multidomain modular type I PKSs using 3-amino-5-hydroxybenzoic acid (AHBA) as the starter unit. Based on the structures of AHBA-derived aromatic moieties, ansamycins are divided into benzenic and naphthalenic groups. Each group can be further classified according to the numbers and types of extender units recruited. So far, only about 200 ansamycins have been reported. By analogy to the biosynthesis of other types of type I polyketides such as macrolactones, we speculate that more ansamycin scaffolds, e.g. hexaketide, dodecaketide and tridecaketide backbones, are waiting for exploitation. In the existing categories, both of benzenic and naphthalenic types have been identified only in the novel ansamycins divergolides (53,54). Moreover, novel ansamycins with unusual extender units have been identified recently (55).

Although ansamycin antibiotics are important secondary metabolites of microorganisms, and rifampicin, the derivative of rifamycin, has been used for more than 50 years in clinic, we don't know much about the molecular mechanisms that regulate their biosynthesis. The current literature reports are limited to the regulation mechanism of the biosynthesis of rifamycin, ansamitocin and geldenamycin. RifO, P, Q in rifamycin biosynthetic gene cluster (rif) may be involved in the regulation of rifamycin biosynthesis,
and rifO may be involved in the regulation of B-factor (56). RifP encodes a transmembrane protein, a member of the DHA14 Drug:H⁺ antiporter family, that pumps antibiotics out of cells using the H⁺ gradient produced by respiration. The antisense gene strategy partially silenced the expression of rifP, which resulted in the reduction of the production of rifamycin B by 70%, confirming the proton pump function of RifP (57). The rifQ gene is highly homologous to the transcriptional suppressor gene actil-ORF1 (58). The latter belongs to the TetR family of transcriptional repressor proteins, indicating that RifQ may be involved in the self-regulation of rifPQ operon. In addition, the protein encoded by rif-orf3 may be involved in the negative regulation of rifamycin biosynthesis. The production of rifamycin B in rif-orf3 knockout mutant increased by about 40% (39). Additionally, gdmR1, gdmR2, orf19 and orf20 may be involved in the regulation of geldanamycin biosynthesis, and orf16 and orf17 may encode resistant genes (60). Gene knockout and complement experiments confirmed that the LuxR family regulators GdmRI and GdmRII positively regulated the transcription of PKS genes, but did not regulate the expression of post-PKS modification genes such as gdmN (61).

Though the biosynthetic mechanism of ansamycins has been studied extensively (62), much is remained elusive about the regulation of their biosynthesis. However, tremendous efforts have been contributed to enhance the production of ansamycins by optimizations of culture media and fermentation conditions (63,64). Indeed, there are ten possible regulatory genes in the ansamitocin biosynthetic gene cluster, which are asm2, asm8, asm18, asm29, asm31, asm34, asm35, asm39, asm40 and asm48, respectively (65). Asm2, Asm29, and Asm34 belong to the TetR family transcriptional regulators. Knockout of the asm2 gene resulted in a nine-fold increase in the production of ansamitocin P3 (66), showing that Asm2 plays a negative regulatory role in the biosynthesis of ansamycins. But surprisingly, overexpression of the asm2 gene also led to 1.3-fold increase in the production of ansamitocin P3 (67). Asm31, Asm39, and Asm40 are sigma factor, sigma factor antagonist and anti-sigma antagonist, respectively. Specifically, Asm39 is a regulatory protein having a histidine kinase-like ATPase C-terminal domain. The constitutive overexpression of the asm39 gene also increased the yield of ansamitocin P3, but the overexpression of either the asm29 or the asm34 gene did not significantly affect the yield of ansamitocin P3 (67). Asm8 and Asm48 belong to LuxR or LAL (large ATP-binding regulator of the LuxR family transcriptional regulators, depletion of the asm8 gene resulted in the complete loss of ansamitocin production (65). Asm18 is homologous to the SARP family proteins. Bioinformatics analysis suggested that Asm18 is a putative transcriptional regulator, but its gene was refractory to amplification from the genomic DNA (gDNA) of Actinosynnema pretiosum (67).

Very recently, the function of the asm18 gene was studied through overexpression in the HGF052 strain (A. pretiosum ATCC 31565Δasm19). The results of transcription and HPLC analysis indicate that asm18 gene encodes a positive transcriptional regulator of the asm gene cluster. Real-time PCR analysis showed that the overexpression of the asm18 gene specifically increased the transcription levels of the genes involved in the biosynthesis of the starter unit 3-amino-5-hydroxybenzoic acid (AHBA) (asm43), polyketide assembly (asmA), post-PKS modification (asm21), as well as the transcription levels of the regulatory gene (asm8). Asm8 is a pathway specific LAL-type activator in ansamitocin biosynthesis. Previous study revealed its positive regulatory role in ansamitocin biosynthesis (65). Further investigation of the chemical constituents of asm18-overexpressed HGF052 strain (HGF052+pJTU824-asm18) led to the isolation of seven ansamitocin derivatives (68). Interestingly, these ansamitocin derivatives include maytansinoid, which is an important component for the production of antibody-maytansinoid conjugates (69).

6. Activation of cryptic ansamycin gene clusters

Previously, through PCR screening of the AHBA synthase genes, we obtained dozens of AHBA synthase gene-positive strains. However, the corresponding ansamycins of most strains were refractory to identification under the conventional laboratory experimental conditions, indicating that the gene clusters containing the AHBA synthase gene were not expressed or just expressed at very low levels (70). S. sp. LZ35 is one of such AHBA-positive strains, and produces two different types of ansamycins, geldanamycins and hygrocins, under the conventional laboratory conditions (71-73). Genome sequence analysis of the LZ35 strain revealed the presence of one more ansamycin gene cluster, namely nam gene cluster, besides these two ones responsible for the biosynthesis of geldanamycins and hygrocins. The pks genes, namA-E within the nam gene cluster, contain eight modules, implying that the nam cluster encodes the biosynthesis of octaketide ansamycins. Bioinformatics analysis of the eight acyltransferase (AT) domains in NamA-E revealed that the possible extender units for the assembly of polyketide chain could be different from those in known octaketide ansamycins, i.e. geldanamycins and ansamitocins, and that the products of this gene cluster could contain a novel scaffold. However, no predicted ansamycins were obtained by large scale fermentation under various conditions. Moreover, reverse transcription (RT) PCR analysis indicated that all the genes examined were expressed poorly in the SR101 strain, a geldanamycin-
nonproducing mutant of the strain LZ35 (73), indicating that the nam gene cluster was cryptic or weakly expressed under the conventional laboratory conditions.

To induce the expression of the nam gene cluster, firstly promoter replacement was carried out by introducing the strong constitutive promoter ermE* at upstream of the pks and amide synthase gene operon in the SR101 strain to generate the SR101namA-ermE* strain. Fermentation and isolation of the SR101namA-ermE* strain resulted in the identification of a tetraketide, a derivative of SY4B under usual conditions. SY4b is one of the major tetraketides accumulated in ΔrifF mutant (rifamycin amide synthase gene deletion) of Amycolatopsis mediterranei S699 (74). These results indicated that promoter replacement only partially activated the nam gene cluster. Further analysis of the nam gene cluster identified nam1 gene encoding a putative positive regulator of the LuxR family proteins. Members of the LuxR family have been previously reported as the activators of polyketide biosynthesis (75, 76). Constitutive overexpression of the nam1 gene was carried out by being cloned into the conjugative and integrative vector pJTU824, placed under the control of the ermE* promoter. The resulting construct was integrated into the chromosome of the SR201 strain (73), which is a mutant derived from the SR101 strain with the deletion of hygrocin biosynthetic genes to create the SR201nam1OE strain. Fermentation of the SR201nam1OE strain led to the isolation of three novel ansamycins, namely neoansamycins A-C (Figure 2) (77), which represents the first successful example of genome mining of ansamycins and illustrates the great potential for exploration of the huge reservoir of cryptic ansamycin biosynthetic gene clusters to obtain novel ansamycins (70, 78).

Phylogenetic analyses showed that most ansamycin biosynthesis gene clusters are flanked by LuxR-type regulators, e.g. geldanamycins (GdmRI, GdmRII) (61), hygrocins (Hgc1) (73), degivolesides (Div8) (79) and ansamitocins (Asm8) (80). LuxR-type regulators usually contain a DNA-binding helix-turn-helix domain but lack γ-butyrolactone-binding or response regulatory domains. These LuxR-type regulators are activators of ansamycin biosynthesis. However, the mechanism of action of these LuxR-type regulators is still poorly explored. The constitutive overexpression of nam1 induced the activation of the cryptic ansamycin gene cluster nam in S. sp. LZ35, further supporting the critical role of LuxR-type regulators in positive regulation of ansamycin biosynthesis. Comparative transpositional analysis with quantitative RT-PCR (qRT-PCR) showed that five genes namA, namK, nam2, nam7 and nam8 involved in the biosynthesis of neoansamycins were upregulated under nam1-overexpressing condition in comparison to the control. The upregulation of transcription of the nam gene cluster, coupled with the isolation of neoansamycins in the SR201nam1OE strain, further confirmed that Nam1 acts as a direct activator of the nam gene cluster. The interaction of Nam1 with the promoter regions of these genes was verified by electrophoretic mobility shift assays (EMSA) performed using purified N-terminal truncated Nam1 protein (Nam1DBD) and the PCR-generated promoter fragments of these nam genes, respectively. Nam1DBD retarded the mobility of all these fragments, suggesting that Nam1 binds specifically to these promoter regions. In order to identify the Nam1 binding site, the promoter region fragments shown to be retarded in EMSAs were studied by DNase I footprinting analysis. Footprinting analyses revealed protected sequences of approximately 20 to 30 nucleotides in target promoters. The comparison of the protected sequences identified the consensus binding sequence and a sequence logo (81) that depicts the binding site. The presence of highly conserved binding sites in these nam genes suggested a strong relevance of this class of transcriptional regulator in the regulation of nam gene expression. Moreover, the knowledge of the LuxR regulators may set the stage for understanding the genetic control of ansamycin antibiotic biosynthesis regulation and provide an effective strategy to discover and improve the yields of these antibiotics.

Additionally, another AHBA synthase gene-positive strain S. sp. XZQH13 contains a cryptic ansamycin biosynthetic gene cluster ast as well. The constitutive overexpression of the astG1 gene, encoding a Large-ATP-binding regulator of the LuxR family, induced the expression of the ast gene cluster including these key biosynthetic genes astB1, astD1 and astF1. This led to the isolation of two ansatrienins hydroxymycotrienin A and thiazinotrienomycin G, further supporting that LuxR-type regulators are pathway-specific positive regulators for the biosynthesis of ansamycins (82). Moreover, overexpression of the LuxR-type regulator gene div8 increases the production of divergolides in the S. sp. W112 strain (79), which further facilitates the isolation of five new ansamycins, namely divergolides O-S (83).

7. Perspectives

With the aid of whole genome sequences of thousands
of bacterial genomes and metagenomes, the fast development of genome mining has dramatically facilitated the automated discovery of promising new natural products, however, mostly ignored their roles in the environments. Though this development has been avoid of duplications in the discovery of natural products incurred by bioactivity-guided fractionation, it is insufficient in function predictions. This gap can be amended by adding bio-filters during genome mining (21). Specifically, microbial libraries for producers of both glycopeptide and ansamycin antibacterial compounds were enriched by screening the self-resistance mechanism of antibiotic producers. Together with screenings for biosynthetic genes resulted in the discovery of a new glycopeptide antibiotic, pekiskomycin (84). The self-resistance guided genome mining approach can be also applied in the discovery of inhibitors of target proteins beyond microorganisms. Panter et al. utilized the pentapeptide repeat protein (PRP)-mediated self-resistance mechanism against topoisomerase inhibitors to search the genome of Pyxidicoccus fallax. A gene cluster was identified adjacent to a predicted PRP (co-clustering), however, no corresponding compounds were reported before their study. Activation in the native host and heterologous expression of this gene cluster resulted in the isolation of pyxidicyclines, a group of new inhibitors of E. coli DNA topoisomerase IV and human DNA topoisomerase I (85). Similarly, the fungal sesquiterpenoid aspterric acid, a potent herbicide, was obtained by the approach of co-clustering self-resistance gene and biosynthetic gene cluster as well (86). Indeed, the self-resistance mechanisms of microorganisms against natural products are diverse (50,87-89) thus that the self-resistance-directed genome mining would be fruitful in the focused discovery of novel bioactive natural products (90,91).

Some natural products are produced by microorganisms evolved from natural selection for acquisition of improved defence against competing deleterious organisms. The production of these defensive secondary metabolites is inducible because they originally serve as responses to environmental challenges. Therefore, the combination of chemical ecology and genome mining can be an integrative approach to the discovery of novel bioactive natural products (92). Much more ambitious, Medema and Fischbach suggested to use globe-wide metagenomes to footprint microbial ecosystems, and then to target these for exhaustive single-cell sequencing, which would facilitate our understanding the ecological roles, and even further the natural roles, of natural products. Ideally, computational analysis can predict the action targets of natural products with the availability of ecological distribution and evolution of biosynthetic gene clusters (93).

Besides ecology, evolution is another critical issue of natural products that should be considered during genome mining. Natural products are produced by secondary metabolism. Secondary metabolism is the response and adaptation mechanisms of organisms to environmental biotic and abiotic stresses, and natural products are the executors of these mechanisms. Based on the evolutionary theory of natural selection, biological traits are derived from the selection of variations in a population. It is beneficial to improve the fitness of organisms to their environments. The mutations that reduce fitness are directly harmful or irreparable, and will disappear from the population. Some of the new properties of variations have no profit or loss, so they belong to neutral variations. Applying these ideas to natural products, it is generally believed that the cost will increase in producing new natural products. If these costs are accompanied by benefits, organisms will be more adaptable. Moreover, mutations of biosynthetic gene clusters are beneficial to fitness but are not lethal to organisms, which exactly fits the nearly neutral theory that the adaptation may be due not to strong selection of rare variants with large effects, but to weak selection of common variants (94-96). Thus, computational analysis of the evolution of the biosynthetic gene clusters in sequences and organizations could help in simulating the evolution of pathways to facilitate genome mining (97).

The phylogeny-guided genome mining represents an approach on the basis of evolutionary biology to the discovery of natural products. This approach was established on the assumption that "a single biosynthetic gene that must have co-evolved with its respective biosynthetic gene cluster could be used as a phylogenetic marker that might represent an evolutionary path of its entire biosynthetic gene cluster" (98). This approach provides a very straightforward means to screen a large number of microbial genomes or metagenomes attributed to the use of a single gene as a molecular marker in search for new biosynthetic gene clusters of interest, which has already resulted in some fruitful findings (98). In particular, with the advance of natural product phylogeny, e.g. the Bayesian analysis of biosynthetic modules and/or domains rather than the sequence of a single gene (99), novel algorithms-promoted integrative pan-genome mining approach will play an important role in both the mining and the synthetic biology of natural products (100).

8. Conclusion

In conclusion, genome mining has brought revolutionary influence on the discovery of novel natural products and the characterization of biosynthetic pathways. The future genome mining would evolve to an integrative approach by the involvement of multiple disciplines to bioactivity-targeted mining of novel natural products.
Acknowledgement
This work was supported in part by the National Natural Science Foundation of China (81530091), and Program for Changjiang Scholars and Innovative Research Team in University (IRT_17R68).

References


(Received November 1, 2018; Revised December 25, 2018; Accepted December 28, 2018)