特约评述

DOI: 10.12211/2096-8280.2021-012

基因组挖掘在天然产物发现中的应用和前景

杨谦[1]，程伯涛[1]，汤志军[1]，刘文[1,2]
（[1] 中国科学院上海有机化学研究所，生命有机化学国家重点实验室，上海 200032；[2] 中国科学院上海有机化学研究所，湖州生物制造中心，浙江 湖州 313000）

摘要：天然产物一直以来都是药物先导化合物的重要来源。在药物发现领域，基因组数据常用来识别潜在的药物靶点或寻找先前被忽视的天然产物的生物合成基因簇。尽管基因组测序发现了微生物和植物中存在大量未开发的化学多样性，然而，仅仅利用传统的分离分析方法获取新的天然产物已经无法满足药物发展的需求。随着基因组时代的到来，数字化的基因组挖掘已经成为天然产物发现的重要组成部分。伴随着高通量测序方法的发展和DNA数据的丰富，各种基因组挖掘方法和工具被开发出来，以指导发现和表征这些天然产物。本文综述了近年来基因组挖掘的网络工具、数据库和方法，着重介绍次级代谢产物生物合成基因簇的挖掘手段，从经典的基因组挖掘到基于抗性基因挖掘、基于系统进化发育的挖掘，并对基因组挖掘在天然产物发现中的地位和前景进行了展望。

关键词：基因组挖掘；天然产物；网络工具；数据库

中图分类号：Q31 **文献标志码：**A

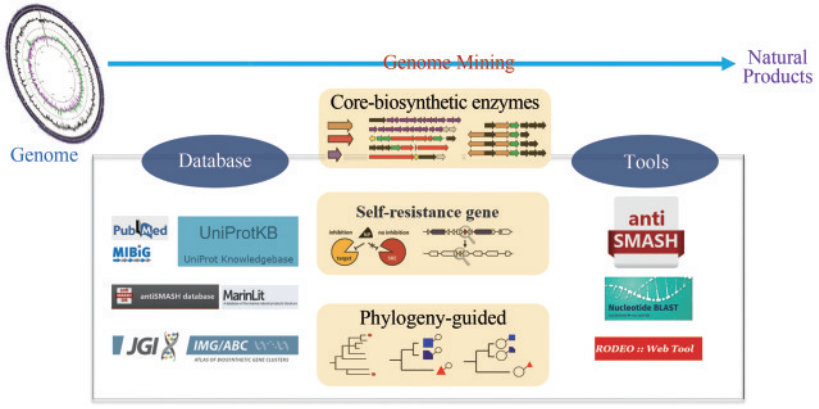
Applications and prospects of genome mining in the discovery of natural products

YANG Qian[1], CHENG Botao[1], TANG Zhijun[1], LIU Wen[1,2]
([1]*State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China*; [2]*Huzhou Center of Bio-Synthetic Innovation, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Huzhou 313000, Zhejiang, China*)

Abstract: Natural products have been an abundant source of leader compounds for new drugs, but traditional isolation and analysis technologies to obtain novel natural products cannot satisfy the requirement for drug discovery. Genomic data have been utilized for identifying potential drug targets, or exploring biosynthesis pathways for natural products that were neglected before. Genome sequencing has unveiled a plethora of undeveloped chemical diversity in microorganisms and plants. From genome sequences, a large amount of information is available, from functional enzymes to conserved patterns/signatures, even potential structures and features that can be interpreted to hunt for new biocatalysts. With the advent of the genomic era, the computational mining of genomes has become an important part in the discovery of novel natural products as drug leads. Meanwhile, the development of high-throughput sequencing

收稿日期：2021-01-27 修回日期：2021-04-05
基金项目：国家重点研发计划（2019YFA0905400）；中国科学院B类先导科技专项（XDB20020200）；王宽诚率先人才计划
引用本文：杨谦，程伯涛，汤志军，刘文. 基因组挖掘在天然产物发现中的应用和前景[J]. 合成生物学，2021，2(5)：697-715
Citation：YANG Qian，CHENG Botao，TANG Zhijun，LIU Wen. Applications and prospects of genome mining in the discovery of natural products[J]. Synthetic Biology Journal，2021，2(5)：697-715

and the establishment of DNA database, genome mining methods and tools have contributed to the discovery and characterization of these natural products. In spite of the diversity of natural products, the biosynthetic rules and thus the biosynthetic machineries for many of these compounds are often remarkably conserved, which is highlighted in the high amino acid sequence similarity of the core biosynthetic enzymes, such as polyketides synthases (PKS), non-ribosomally peptides synthetases (NRPS), and many others. Besides, most of natural products are considered to be produced by the host to kill or limit the growth of competitors through the inhibition or inactivation of essential housekeeping enzymes. Therefore, accumulating knowledge on the self-resistance mechanisms, for instance, mining for SRE (self-resistance enzyme), have promoted research on natural products. Moreover, a phylogeny-guided mining approach provides a method to quickly screen a large number of microbial genomes or metagenomes to detect new biosynthetic gene clusters of interest, and many web tools and databases have been developed and utilized by researchers to mine for key enzymes. This paper reviews recent advances in the genome mining tools, databases and approaches, with a focus on the ways of mining biosynthetic gene clusters (BGCs) of natural products, from classical genome mining to resistance-based and phylogeny-guided mining, and also include a short overview on status and perspective in the discovery of novel natural products.



Keywords: genome mining; natural products; web tools; database

自然界作为活性天然产物的重要来源孕育了数以万计的生命有机体。在过去的几十年里，绝大多数抗癌、抗感染和抗菌药物都源于生命体所产生的天然产物及其衍生物，如青霉素、洛伐他汀、紫杉醇[1-2]。其中，许多来源于土壤、海洋及特殊环境的微生物由于能够产生一系列活性显著且具有成药潜力的天然产物而备受关注，这些化合物的发现也为新药创制开辟了新的思路。然而，人类疾病谱的变化以及多药耐药等一系列问题的出现，使得开发新的药物成为人类健康的迫切需求。基于活性导向药物发现的方法虽然可以从植物、动物及微生物中分离获得具有生物活性的小分子，但是这些筛选方法不仅耗时耗力，而且不能避免重复性、盲目性以及低效率等弊端。

基因组时代的到来为我们提供了来源于不同有机体数以万计的DNA（deoxyribo nucleic acid，脱氧核糖核酸）序列，这些数据不仅能够为生物学各个分支学科提供丰富的物质基础，同时也为天然药物的发现带来了新的曙光。基因组数据不仅可以用来识别潜在的药物靶标[3]，还能用于寻找先前被忽视的次级代谢产物的生物合成途径[4]，定向发现新的天然产物药物。每一个旨在预测生理或代谢特性的生物信息学研究都可以被认为是基因组挖掘（genome mining）。然而，在与天然产物相关的文献中，“genome mining”经常被赋予更特殊的意义，它不再局限于通过计算模拟的方法

检测生物合成基因，还涉及到它们的功能研究，最终阐明相关的化学机制。随着基因组数据的丰富，次级代谢产物的基因簇不再匮乏，更大的挑战转向了如何高效快速地锁定具有挖掘潜力的生物合成基因簇（biosynthetic gene clusters，BGCs），从而快速地获得药物实体分子。在生物信息学发展的同时，许多专业的网络工具也被研究人员运用到基因组数据的挖掘过程中，目前已有许多综述进行了报道[5-13]。本文综述了基因组挖掘在天然产物发现中的运用，包括最新的生物信息学工具、常用的各类数据库以及主要的挖掘方法，并对基因组挖掘在药物发现和多个学科领域中的影响和发展进行了展望。

1 基因组挖掘过程中的数据库和网络工具

数字革命正在改变人们储存、传播和使用信息的方式。随着关联数据、机器学习和大规模网络推理等新技术的出现，天然产物研究领域开始实现数字化实验数据的实时共享和大规模分析。数据库在这方面发挥了关键作用，因为它们允许对基本和高级应用程序的数据进行系统的注释和存储[14]。

1.1 天然产物数据库

2020年，Maria Sorokina等整理了2000年以来所有的天然产物数据库，到目前正在使用的数据库超过了120个，其中只有50个可以免费访问[15]。在这些数据库中，有3个在微生物领域运用最为广泛，即NPASS、StreptomeDB和The Natural Products Atlas。其余常用的数据库还包括Dictionary of Natural Products（DNP）[16]、PubMed[17]、MarinLit、AntiBase、KNApSAcK[18]、Norine[19]（非核糖体肽专门的数据库）和MacrolactoneDB[20]等，这些数据库包含了来源于植物、海洋微生物、真菌及细菌等生命有机体产生的天然产物。

NPASS是2018年发展起来的一个数据库[21]，旨在提供天然产物的来源及其生物活性。它包含了部分来自植物、无脊椎动物和微生物天然产物的化学结构，共含有超过35 000种化合物，其中大约9000种来源于微生物。StreptomeDB是一个链霉菌属天然产物的专门数据库[22]。在2020年的更新信息中，该数据库包含了7100多个化合物的来源、生物活性及其光谱信息。The Natural Products Atlas[23]是2019年建立的一个新的数据库，它提供了所有微生物产生的天然产物衍生物的化学结构。当前该数据库包含了超过25 500个化合物，它具有一个特殊的检索链接，能够双向连接到另外两个天然产物资源库：一是生物合成基因簇的MIBiG（Minimum Information about a Biosynthetic Gene cluster）[24]数据库；另一个是天然产物质谱数据的GNPS（Global Natural Products Social Molecular Networking）[25]数据库。Dictionary of Natural Products是一个非开源数据库，主要收录天然产物的来源、物理特性及其生物学数据，目前已经收录了超过29万个条目。MarinLit是一个海洋天然产物的文献数据库，包含超过35 000个化合物的化学结构、分类及其全合成数据。它是目前海洋天然产物最新和最全面的数据库。值得一提的是，目前世界上最大的两个非开放的天然产物数据库：Scifinder和Reaxys。这两个平台包括了来自天然产物文献的大多数天然产物及其衍生物、合成中间体。

1.2 生物合成基因簇数据库

自2000年以来，越来越多的基因组数据被报道，而在NCBI GenBank[26]中能够找到“基因-结构”相对应的数据屈指可数。为了解决这一问题，研究者开发了一系列专门的生物合成基因簇数据库，例如ClusterMine360[27]、DoBISCUIT（Database of BIoSynthesis clusters CUrated and InTegrated）[28]、MIBiG 2.0[24]、IMG-ABC[29]、antiSMASH Database[30]和Recombinant ClustScan Database[31]。ClusterMine360作为早期的生物合成基因簇分析平台，将那些通过实验证实的生物合成基因簇与其对应的化合物进行了链接，主要聚焦于非核糖体肽（non-ribosomal peptide，NRP）和聚酮（polyketide，PK）类化合物，它包含了超过300个天然产物生物合成基因簇的信息。2015年，通过150多名天然产物科学家的

共同努力发布了“生物合成基因簇的最小信息库”（Minimum Information about a Biosynthetic Gene Cluster，MIBiG），对已被实验证实天然产物的生物合成基因簇进行了人工注释[32]。利用联合基因研究所（Joint Genome Institute，JGI）的细菌基因组平台发布的IMG/M[29]和IMG-ABC[29]，旨在发展成为一个最全面的细菌基因组数据库，它包含已知天然产物的生物合成基因簇（有些信息间接地来源于MIBiG）的信息，并且能够通过计算模拟预测未知生物合成基因簇的功能。到目前为止，该数据库包含了来源于antiSMASH和ClusterFinder算法模拟获得的超过100万个生物合成基因簇。由于JGI的数据使用限制，部分数据目前处于未公开状态。同样由JGI开发的真菌基因组门户MycoCosm[33]，包含超过1000个真菌基因组信息，它通过提供交互式网络工具，支持真菌基因组序列和其他“组学”（omics）数据的整合、分析和共享。2016年，antiSMASH的开发团队发布了antiSMASH database[30-34]（antiSMASH-DB），作为antiSMASH运行的中央存储库。与IMG-ABC不同的是，antiSMASH-DB旨在提供一个有限的、复杂的假定生物合成基因簇列表，这些生物合成基因簇来自最高质量的细菌基因组。在2018年发布的第二版中，antiSMASH-DB包含了来源于24 000多个细菌基因组的152 000多个生物合成基因簇[35]。

1.3 蛋白家族数据库

蛋白质通常由一个或多个功能区域组成，识别这些功能区域有助于预测未知蛋白的功能。UniProtKB[36]是目前世界上最大的蛋白数据库，囊括了超过2亿个非重复的条目。它分为两个部分：UniProtKB/SwissProt和UniProtKB/TrEMBL。其中，UniProtKB/SwissProt带有功能性的注释，这些注释来源于各类文献中已经被人工核实的数据。截至2020年6月，SwissProt收录了563 972个条目，由于人工注释并不能做到面面俱到，因此它收录的功能并不是全面的，但是SwissProt能够接受使用者输入注释数据，从而达到数据库的实时更新。在UniProtKB/TrEMBL中，基于序列同源性的分析会进行自动分配注释，系统会自动将满足条件的注释从储存序列转移到假定的同源序列中[37]。

目前，常用的蛋白家族数据库包括Pfam[38]和InterPro[39]。Pfam[38-40]是注释信息最为全面的蛋白家族数据库，每个家族都由多序列比对和隐马尔可夫模型（HMMs）表示。在最新发布的版本Pfam 33.1中，定义了18 259个同源家族，有70%的条目与UniProtKB密切相关。InterPro是一个更大的蛋白家族数据库。截至2020年11月，该数据库定义了38 345个条目，包括3326个同源超家族、22 618个家族、11 162个功能域。但是，并不是所有的序列都能被Pfam和InterPro所包含，因此，在蛋白质领域存在着大部分尚未注释的基因组“暗物质”，它们可能具有某些特殊的功能[5]。为此，生物学家开发出了一系列能够有效挖掘蛋白和基因组数据的工具，称为“基因组酶学计算机模拟工具”[5]。

1.4 识别生物合成基因簇的网络工具

自从第一个链霉菌基因组被测序[41]，基因组挖掘迅速成为发现天然产物的一种重要方式，人们发现神秘的生物合成基因簇为新颖天然产物的发现开启了新的篇章。基因组挖掘利用遗传/基因组信息来评估微生物产生新化合物的遗传潜力，通过运用多种生物信息学工具在众多的基因序列中识别具有潜在价值的生物合成基因簇，并根据基因簇中的信息来预测其产物，最终阐明生物合成过程。

序列相似性搜索是一项非常重要的生物信息学任务。基于局部比对搜索工具BLAST（Basic Local Alignment Search Tool）[42]和利用隐马尔可夫模型（hidden Markov model）进行蛋白序列分析的网络工具HMMer[43]是目前最简单的序列比对工具，它们使用人工构建的基因列表作为查询序列，对未知蛋白进行序列比对从而初步预测其功能。此外，DIAMOND[44]是一种基于双索引的开源算法，速度比BLASTx快20 000倍，但灵敏度与其不相上下。如今，这些分析方法已经变得越来越复杂，常用的分析工具包括：ClustScan（Cluster

Scanner）[45]、CLUSEAN（CLUster SEquence ANalyzer）[46]、np.searcher[47]、SMURF[48]和antiSMASH[49-50]。ClustScan是一个用于模块化生物合成基因簇的半自动注释和新型化学结构的计算机模拟预测的集成程序包。该程序包用于快速、半自动地对编码模块化生物合成酶的DNA序列进行注释，包括聚酮合酶（PKS）、非核糖体肽合成酶（NRPS）和聚酮-非核糖体杂合酶（PKS/NRPS）。但由于版权限制目前只能提供用户30天的试用期，属于半开放的程序包。CLUSEAN是一个开放式的自动分析细菌次级代谢产物生物合成基因簇的计算机框架程序。它集成了标准的分析工具，如BLAST和HMMer，以及能够识别非核糖体肽合成酶（NRPS）/Ⅰ型聚酮合酶（TIPKS）功能域和基序的特定工具，并且能够预测NRPS的特异性。

为了促进真菌基因组中次级代谢产物生物合成基因簇的系统定位，Natalie D. Fedorova等开发了一个基于网络的软件工具——Secondary Metabolite Unique Regions Finder（SMURF）[48]，它基于真菌次级代谢产物生物合成途径的三个特征：①存在骨架基因；②成簇出现；③含有特征蛋白的结构域，对真菌基因组中的生物合成基因簇进行预测和归类。2011年，Eriko Takano等开发一个强大的网络工具，用于细菌和真菌基因组序列中次级代谢物生物合成基因簇的快速识别、注释和分析，并命名为antiSMASH（antibiotics & Secondary Metabolite Analysis Shell）[49]。直至2019年，antiSMASH 5.0[50]增加了编码酰基氨基酸、β-内酯、真菌RiPPs等多种类型基因簇的检测规则，尤其是对于编码Ⅱ型PKS生物合成基因簇提供了更多详细的预测，同时在网页运行方面也缩短了一些时间。

近几年一些新颖的分析工具相继被开发使用，它们能够解决上述算法中的缺陷：在检测已知基因簇方面具有高特异性，但是并不会识别未知的生物合成基因簇。从基因组中预测未知类别基因簇应该是最具优先级的，因为这些基因簇可能编码了全新骨架的分子[51]。根据文献报道，目前实验室可培养的微生物只占总数的1%[52]，而这1%的微生物就含有超过200万株细菌或真菌（http://www.wfcc.info/ccinfo/），这意味着含有数量级的未被识别的生物合成基因簇有待开发和挖掘。这些基因簇被人们统称为“微生物的暗物质”，它们可能存在于未被开发的菌株中，也可能存在于像大肠杆菌这样被研究成熟的微生物中。这就需要运用更为复杂的算法提供强有力的检索能力来识别“暗物质”，从而成功地获取全新的天然产物分子。

目前开发了3个研究策略：①ClusterFinder[53]算法，它首先识别基因组序列中可能的基因编码区域，利用Pfam数据库对编码区域进行蛋白功能域注释，然后依据Pfam数据库中的蛋白功能域在训练集生物合成基因簇中出现的频率，利用HMM将其设定为BGC或者non-BGC状态。ClusterFinder能够识别出富含BGC状态Pfam数据库功能域的基因组区域。这种策略能够发现新类型的基因簇，因为不同分子的生物合成途径往往利用相同家族的酶，如氧化还原酶、甲基转移酶、CoA连接酶和P450氧化酶[53]。②基于所有次级代谢酶都是初级代谢酶同源物这个观点发展了EvoMining[54]方法，通过检测基因组中“额外”的代谢酶，使用系统发育分析来识别进化上具有明显差异的序列，并对其上下游基因进行功能分析，从而发现新的生物合成基因簇。③使用大规模的基因组序列比对。首先利用BLASTp寻找不同基因组中的同源基因，从同源基因出发通过局部比对识别种子区域（seed regions），对种子区域进行扩张，锁定基因簇边界并进行共线性分析（synteny analysis），最终寻找到可能编码新颖次级代谢产物的基因簇[55]。这三种策略的综合运用可能成为未来识别生物合成基因簇最有效的方法[7]。

核糖体合成和翻译后修饰肽（RiPPs）是从基因编码的前体肽衍生而来的一类天然产物，由于不同类别前体肽缺乏共同的序列特征，因此通过计算识别其生物合成基因簇一直是极具挑战的任务。最近开发了几种新的算法，专门进行RiPPs的生物合成基因簇的挖掘。Andrew W. Truman等开发了一种用于识别不同家族RiPP前体肽工具RiPPER，运用该方法在放线菌中找到了新的含有硫酰胺结构的RiPPs[56]。许多RiPPs后修饰的发生

依赖于一个称为RiPP识别元件（RRE）的蛋白结构区域。RRE与前导肽（leader peptide）特异性结合，并引导翻译后修饰酶作用于核心肽（core peptide）。Douglas A. Mitchell等开发了一种基因组挖掘的工具RRE-Finder[57]，它从UniProtKB蛋白数据库中调取25 000条高可信度的RRE蛋白序列作为样本数据库，进一步识别基因组中可能包含RRE序列的生物合成基因簇。此外，还有一些新的挖掘工具也被开发出来，例如DeepRiPP[58]和RODEO（Rapid ORF Description and Evaluation Online）[59]。基因组挖掘过程中常用的数据库及网络工具见表1。

表1 基因组挖掘的数据库及网络工具

Tab. 1 Database and web tools of genome mining

数据库或web工具	网址（URL）	参考文献
天然产物数据库		
Dictionary of Natural Products（DNP）	http://dnp.chemnetbase.com	[16]
The Natural Products Atlas	https://www.npatlas.org	[23]
PubMed	https://pubmed.ncbi.nlm.nih.gov/	
NPASS	http://bidd2.nus.edu.sg/NPASS	[21]
StreptomeDB	http://132.230.56.4/streptomedb2/	[20]
MarinLit	http://pubs.rsc.org/marinlit/	
AntiBase	https://sciencesolutions.wiley.com	
KNApSAcK	http://kanaya.naist.jp/KNApSAcK/	[18]
Norine	https://bioinfo.lifl.fr/norine/	[19]
MacrolactoneDB	https://macrolact.collaborationspharma.com/	[20]
ChEBI	http://www.ebi.ac.uk/chebi/	[60]
ChEMBl	https://www.ebi.ac.uk/chembl/	[61]
ChemSpider	http://www.chemspider.com/	[62]
COCONUT	https://doi.org/10.5281/zenod	[15]
生物合成基因簇数据库		
ClusterMine360	http://www.clustermine360.ca/	[27]
DoBISCUIT	http://www.bio.nite.go.jp/pks/	[28]
MIBiG	https://mibig.secondarymetabolites.org/	[24]
IMG-ABC	https://img.jgi.doe.gov/cgi-bin/abc/main.cgi	[29]
antiSMASH Database	https://antismash.secondarymetabolites.org/	[30]
ClustScan Database	http://csdb.bioserv.pbf.hr/csdb/	[45]
BiG-FAM	https://bigfam.bioinformatics.nl/	[63]
蛋白家族数据库		
UniProtKB	https://www.uniprot.org/	[36]
Pfam	http://pfam.xfam.org/	[40]
InterPro	http://www.ebi.ac.uk/interpro/	[39]
识别生物合成基因簇的网络工具		
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi	[42]
HMMer	http://hmmer.org/	[43]
ClustScan	http://bioserv.pbf.hr/cms/	[31]
np.searcher	http://dna.sherman.lsi.umich.edu/	[47]
SMURF	http://jcvi.org/smurf/index.php	[48]
antiSMASH	http://antismash.secondarymetabolites.org	[49-50]
ClusterFinder	https://github.com/petercim/ClusterFinder	[53]
RODEO	http://rodeo.scs.illinois.edu/	[59]

2 基因组挖掘在天然产物发现中的应用

“基因组挖掘”，几乎与每一个生物信息学研究相关联，它可以用于检测生物活性天然产物的生物合成途径。对天然产物研究领域而言，基因组挖掘就是在没有化学结构的前提下，基于遗传信息来预测和分离活性天然产物。根据挖掘对象的不同，可以大致分为基于核心骨架酶的挖掘、基于抗性基因的挖掘以及基于系统进化的挖掘。

2.1 基于编码核心骨架的酶进行挖掘

以编码合成核心骨架的酶出发，挖掘具有特定结构片段的天然产物，是一种经典的基因组挖掘方法。尽管次级代谢产物的结构多种多样，但是同一类型代谢产物的生源途径往往是非常保守的，这是由于许多核心骨架生物合成的酶在序列上具有高度的相似性。如聚酮类（polyketides）、非核糖体肽类（non-ribosomal peptides）以及氨基糖苷类（aminoglycosides）。利用天然产物结构与其对应的生物合成基因一一对应的关系，在基因层面发现含有特定结构片段的天然产物，指导新化合物的发现。

烯二炔类抗生素是迄今为止发现的抗肿瘤活性最高的天然化合物[64]（图1），其活性中心是双键偶联两个炔键构成的烯二炔核心结构，目前已有20余例烯二炔天然产物陆续被报道，虽然它们的核心环不同（九元环或者十元环），但是核心的烯二炔单元却由相同的生物合成逻辑合成，由包含编码特殊Ⅰ型聚酮合成酶PKSE、硫酯水解酶TE和3个未知功能蛋白在内的5个连续基因组成的基因盒催化完成[65-66]。聚酮合成酶PKSE重复使用7次，完成结构独特的九元（C-1027）或十元（Calicheamicin）烯二炔核心结构的不饱和聚酮前体的合成，再由3个未知功能的酶以及TE催化完成核心烯二炔单元的合成。为了挖掘更多的烯二炔类天然产物，Shen Ben研究组以*PksE*核心基因为探针，对4889个已经测序的微生物基因组分析，又找到51个基因组中含有合成烯二炔结构特征的基因盒[67]；此外，他们还基于实时PCR技术，开发了基于核心基因盒快速分析菌株是否含有合成烯二炔的基因簇的高通量方法，从3000株菌株中找到81株具有烯二炔生物合成基因簇[68]。以上结果表明，虽然目前发现的烯二炔类天然产物很少，

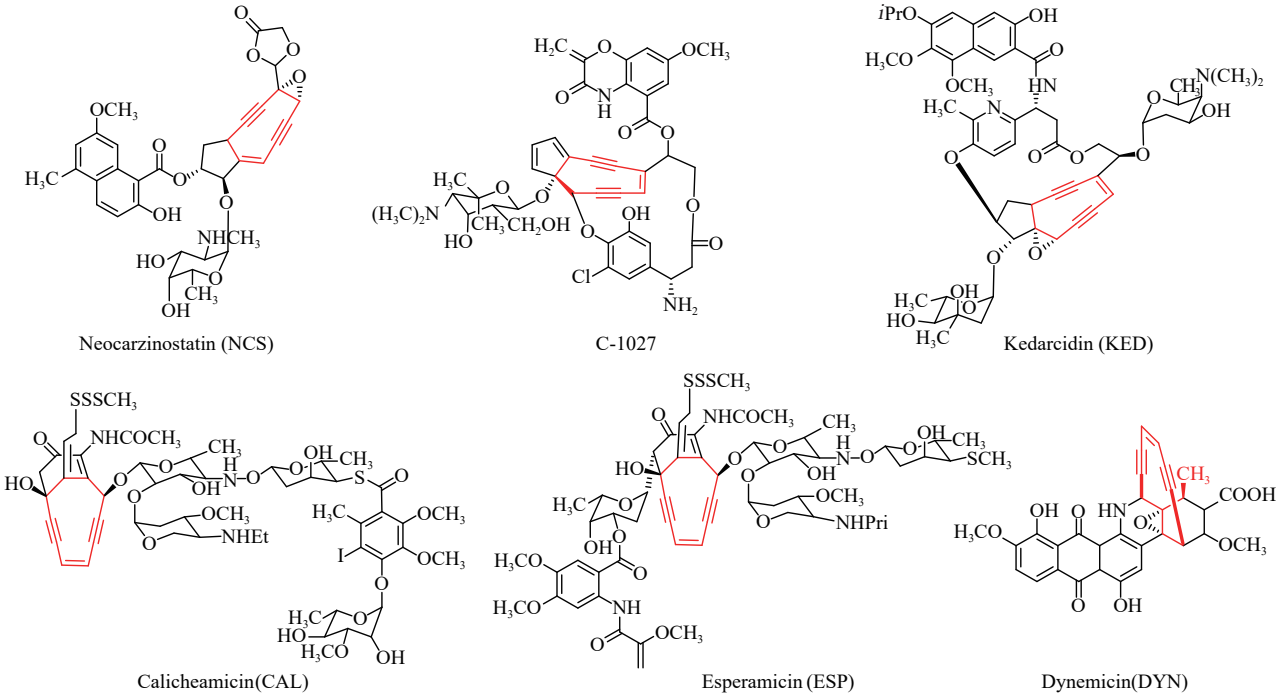


图1 代表性的烯二炔类化合物

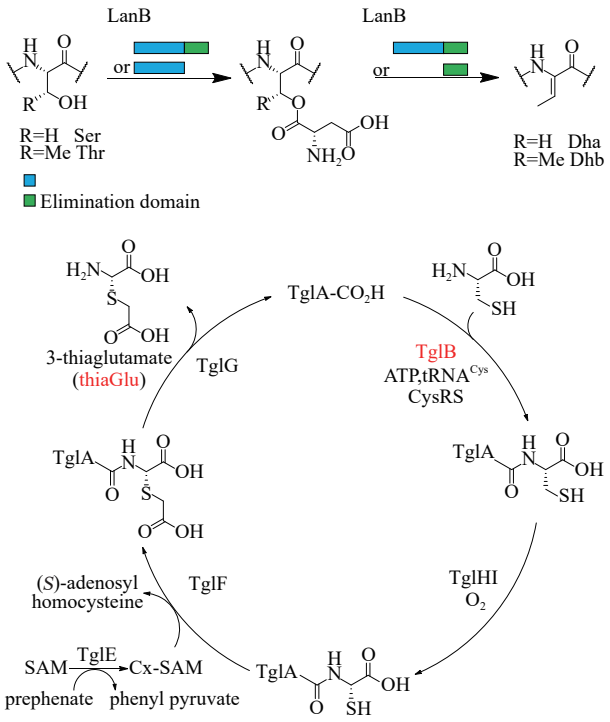
Fig. 1 Representative compounds of Enediyne

但是大自然有巨大的潜力合成更多这类高活性化合物。随着沉默基因激活技术的成熟、异源表达体系的完善、发酵分离技术的提高，有望利用基因组挖掘的方法，分离得到更多、活性更优的烯二炔类天然产物。

脱水氨基酸是羊毛硫肽和硫肽类天然产物共同具备的特征结构片段。Ⅰ型羊毛硫肽和硫肽的脱水氨基酸通过相同的化学机制引入[69-70]。LanB蛋白的谷氨酰化结构域催化前体肽中丝氨酸/苏氨酸侧链羟基的谷氨酰活化，谷氨酸消除结构域催化谷氨酸离去形成脱水氨基酸（图2）。Van der Donk研究组对超过100 000个细菌基因组进行LanB蛋白搜索，寻找到超过600个与LanB谷氨酰化结构域同源的基因，包含这些基因的基因簇或者基因组不包括LanB谷氨酸消除结构域同源基因[71]。他们利用来源于*Pseudomonas syringae* pv. *maculicola* ES4326的*tgl*基因簇为研究对象，揭示了一类新的核糖体肽分子pearlin的生物合成过程。*tgl*簇中包含LanB同源蛋白TglB，其催化前体肽TglA的羧基端半胱氨酰化，在前体肽上实现一分子半胱氨酸的延伸。在整个生物合成过程中，不涉及对前体肽TglA的额外修饰，前体肽仅作为骨架，接受后修饰酶识别，承载修饰对象半胱氨酸。最终，半胱氨酸被转化为thia-Glu成熟分子。

polytheonamide（图3）是一种具备高细胞毒性的核糖体肽类天然产物，成熟分子中含有DL-交替的氨基酸序列和AsmX5Asm的天冬酰胺*N*-甲基化基序，保证成熟分子形成可插入细胞膜的稳定β-螺旋构象[72]。D构型非天然氨基酸的引入由依赖于*S*-腺苷甲硫氨酸的PoyD蛋白负责[73]，天冬酰胺的侧链*N*-甲基化则由*N*-甲基化酶PoyE催化形成[74]。Jörn Piel研究组以polytheonamide合成途径的前体肽基因*poyA*、异构酶基因*poyD*、*N*-甲基化酶*poyE*作为样本序列，分别对非冗余蛋白序列数据库进行BLASTp搜索，集合同时含有三者同源序列的基因组，挖掘到*aer*基因簇[75]。该簇导向发现了polytheonamide类似结构终产物aeronamide A（图3），其同样具备高细胞活性，针对HeLa细胞的IC_{50}值为1.48 nmol/L。

蛋白功能总是处在不断进化的过程中，尽管来源于相同的祖先序列，在经历如基因复制、水平基因转移等生理过程后，基因的功能趋向差异



TglA = MGQPNVQSVENQQASGDVKDLENTPQATEEALFEEFDLDDIEVIESKVFA

图2 LanB蛋白的催化机制及pearlin的生物合成过程

Fig. 2 Catalytic mechanism of LanB and the biosynthesis of pearlin

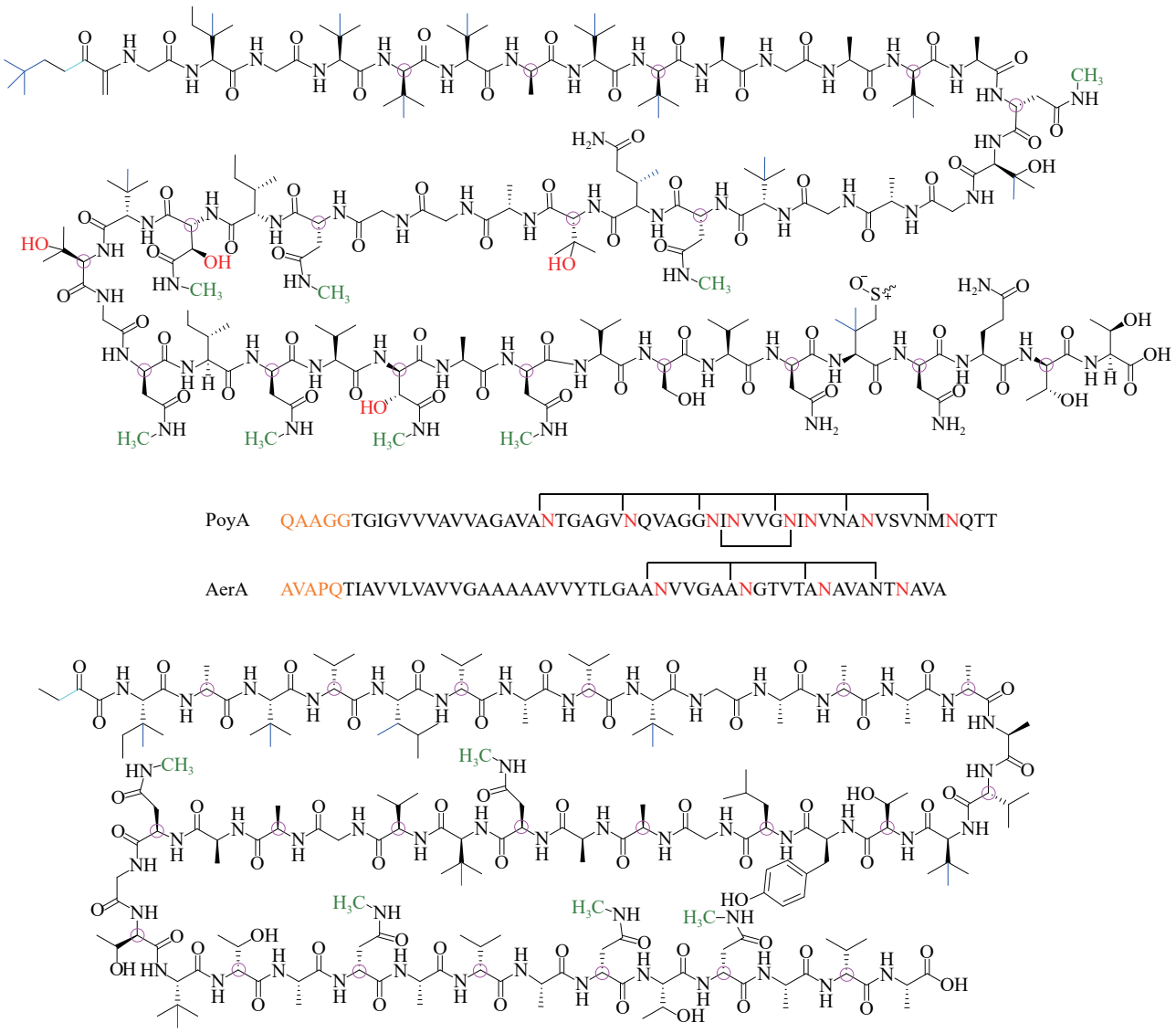


图3 polytheonamide和aeronamide A的化学结构

Fig. 3 Chemical structures of polytheonamide and aeronamide A

化。因此，具备一定序列同源性的蛋白可能存在催化功能上的差异性。蛋白功能的差异导致天然产物结构的差异。从催化特定反应的蛋白出发，建立其与同源蛋白之间的进化关系，可能寻找到催化差异反应的同源蛋白，定位到具备新结构的天然产物，表现出相似或者差异的生理功能。

自由基SAM酶（rSAM，图4）普遍存在于核糖体肽合成途径中，其利用［4Fe-4S］簇还原性裂解*S*-腺苷甲硫氨酸生成5′-脱氧腺苷自由基[76]。该自由基进一步从底物中提取氢原子，从而引发不同的反应。Sactipeptide分子中存在特征性的S—C_α硫醚键，该键由具有额外的C端［4Fe-4S］簇结合基序（SPASM）的自由基SAM酶催化形成[77]。从6个已知的催化前体肽S—C_α键形成的rSAM酶出发，Douglas A. Mitchell研究组利用PSI-BLAST建立蛋白序列文库，并利用RODEO注释备选蛋白本地基因组序列，对潜在的前体肽序列进行打分，通过该流程，他们极大地扩展了Sactipeptides类化合物的序列多样性[78]。

不仅如此，通过对获得的蛋白文库进行序列相似性网络分析（sequence similarity networks，SSN），他们发现与QhpD蛋白同源性较高的rSAM酶催化S—C_β和S—C_γ的形成，不同于已知的S—C_α硫醚键。这一发现拓宽了rSAM酶催化形成的硫醚

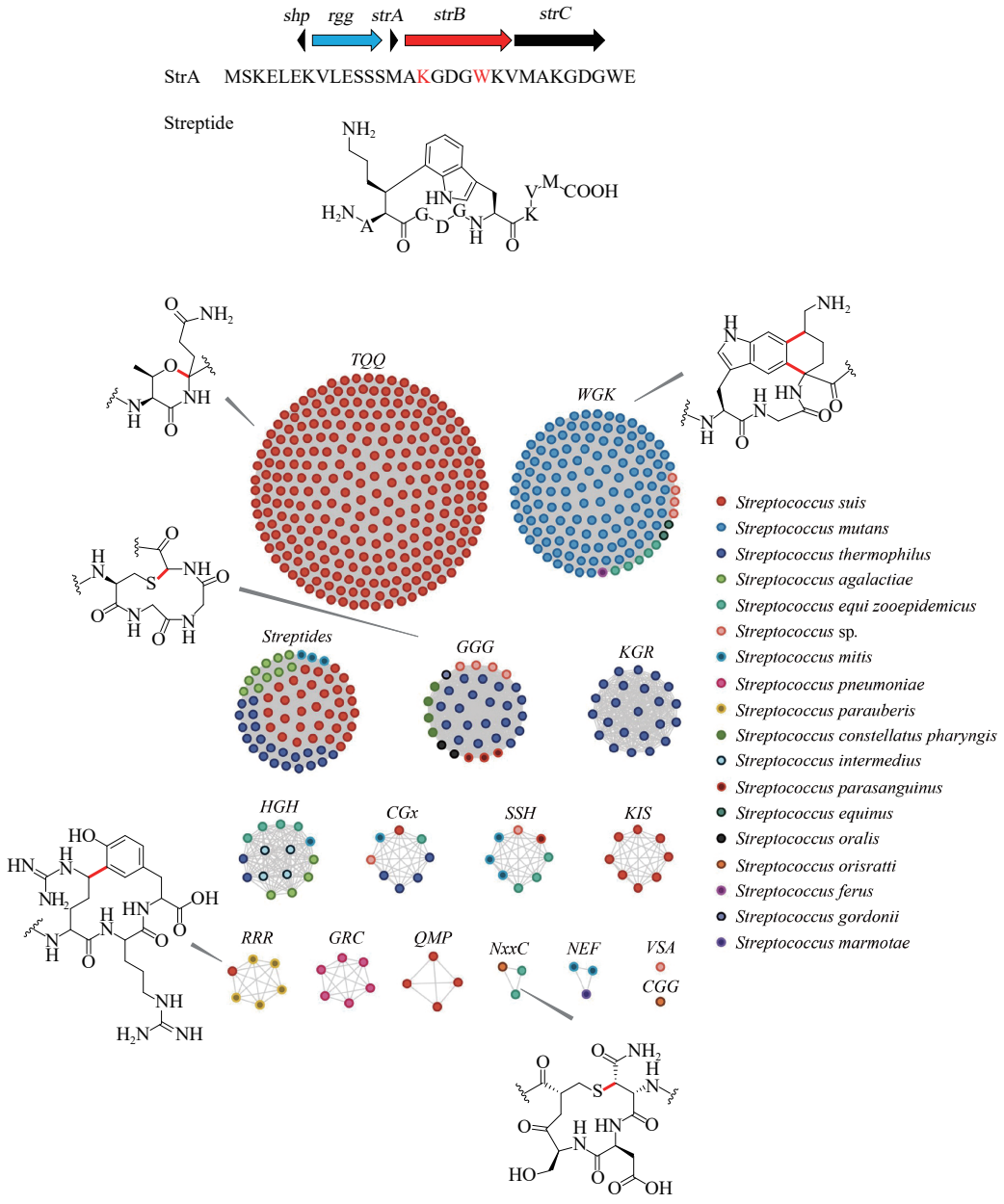


图4 rSAM酶催化形成的产物结构多样性

Fig. 4 Diversity of chemical structures developed by the catalysis of rSAM enzymes

结构多样性。Streptide是一类由链球菌产生的长度为9个氨基酸的核糖体肽，具有特征的赖氨酸C_β和色氨酸吲哚C7位碳碳键偶联结构，该结构由rSAM蛋白StrB催化形成[79]。Streptide的产生受到簇内*shp/rgg*群集感应系统的调控[80]。为了在链球菌中获得更多类似的受到群集感应系统调控的核糖体肽类天然产物，Mohammad R. Seyedsayamdost研究组对集合微生物基因组和微生物组[Integrated Microbial Genomes and Microbiomes（IMG/M） System]上的2875个链球菌基因组进行了分析[81]。利用系统中的Function Profile tool进行搜索，列出所有包含Pfam04055（rSAM蛋白）或者TIGR01716（Rgg/GadR/MutR family）的基因组。最终获得592个同时包含rSAM和Rgg蛋白的潜在基因簇。对所有的rSAM蛋白建立序列相似性网络，并对相应的前体肽生成序列标识图（sequence logo），以确定前体肽的保守序列区域。利用这种方法，他们获得了一系列在*shp/rgg*群集

感应系统调控下可能产生的核糖体肽类产物。它们同样由rSAM蛋白修饰，产生多种类型的化学键，包括碳碳键[81-82]、碳氧键[83]、碳硫键[84-85]。这一发现极大地拓展了rSAM在核糖体肽后修饰中催化形成的结构类型及其酶学功能。

2.2 基于抗性基因的挖掘

活性导向天然产物的发现一直是高通量筛选活性化合物的重要方法。近年来，迅速发展的基因组测序技术使得天然产物的发现发生了革命性的变化。这种以基因组扫描为基础发现天然产物的策略已经成功地发现了许多新颖的代谢产物，并通过大量实验证实了这些天然产物能够极大地增加其化学结构的多样性[86]。尽管这些挖掘方法能够发现独特的生物合成酶和特异的化学物种，但在生物活性方面却没有一个明确的目标，如何利用基因组数据来预测天然产物生物活性成为基因组挖掘的一个热点。为了避免被代谢产物所误伤，微生物在产生活性天然产物的同时进化出了能够抵抗其毒性的基因，使其能够在产生防御机制的同时完整地保存自己。因此，基于抗性基因的挖掘，不仅能够发现结构多样的天然产物，而且能够预测其潜在的生物活性及其作用靶点，为新颖药物的发现提供强有力的研究基础。宿主的抗性或者自我保护机制主要包括以下几种（图5）：其一，外排泵（主动运输代谢产物到细胞外）；其二，对天然产物本身进行修饰从而防御其带来的伤害；其三，修饰宿主内部的管家酶（housekeeping enzyme）来避免天然产物的抑制作用[87-89]。

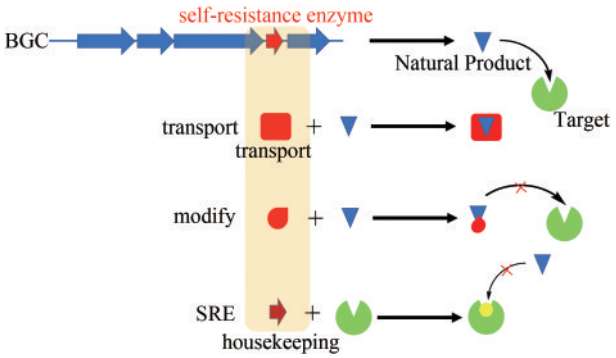


图5 宿主的自抗机制[90]

Fig. 5 Self-resistance mechanism of the host[90]

自然界用于自我保护的另一种策略是编码一个功能等价的自抗性酶（self-resistance enzyme，SRE），它是管家酶的变体。自抗性酶在序列上与管家酶高度相似，它不仅拥有管家酶的功能，同时还能抵御代谢物对宿主的伤害。SRE往往与天然产物生物合成基因成簇存在，也与天然产物生物合成基因同时转录。因此，利用SRE的序列相似性挖掘策略能够快速地定位细菌和真菌天然产物的生物合成基因簇[90-91]。DNA的复制是一个基本的生命过程。然而，这种生命过程在细菌和真菌中却不尽相同。由于这一过程在原核生物中是高度保守的，因此抑制细菌中DNA的复制就成为抗生素筛选的一个理想靶点。G. M. Savage等从葡萄球菌中首次发现了能够抑制DNA复制的抗生素novobiocin，研究表明该化合物的作用靶点是一个DNA旋回酶（DNA gyrase），它属于Ⅱ型拓扑异构酶的一个亚型[92]。其生物合成研究显示，该化合物的生物合成基因簇中存在的*gyrB*基因编码一种对其不敏感的管家DNA旋回酶的变种[93-94]。由于DNA的复制在原核生物与真核生物之间的差异，寻找共同的、具有普适性的抗性基因挖掘策略成为微生物抗生素发现的关键。回溯到生物合成基因簇，参与蛋白质生物合成的酶是开发抗生素的经典靶点。在蛋白质生物合成过程中，转运RNA（tRNA）优先被20个氨基酰化-tRNA合成酶（aminoacyl-tRNA synthetases，aaRSs）编码的同源氨基酸进行酰化。有几个重要的天然产物以此为靶点被挖掘，如mupirocin[95]、thiomarinol A[96]和borrelidin[97]（图6）。在这些天然抑制剂中，mupirocin被FDA批准用于治疗皮肤感染性疾病脓疱疮。

许多参与脂类合成和降解的酶都是有机体所必需的，大部分天然产物的生物合成基因簇以脂肪酸生物合成路径编码的SRE为靶标来实现自抗。来源于真菌最为著名的天然产物洛伐他汀（lovastatin），是一种被FDA批准治疗高胆固醇的药物，它针对的是甲羟戊酸途径限速步骤中的3-羟基-3-甲基戊二酰辅酶A还原酶（HMGR）[98]。在土曲霉中，lovastatin由*lov*生物合成基因簇编码合成，推测该化合物可能是为了对抗真菌中其他的甾醇生物合成途径而产生。在其基因簇中出现

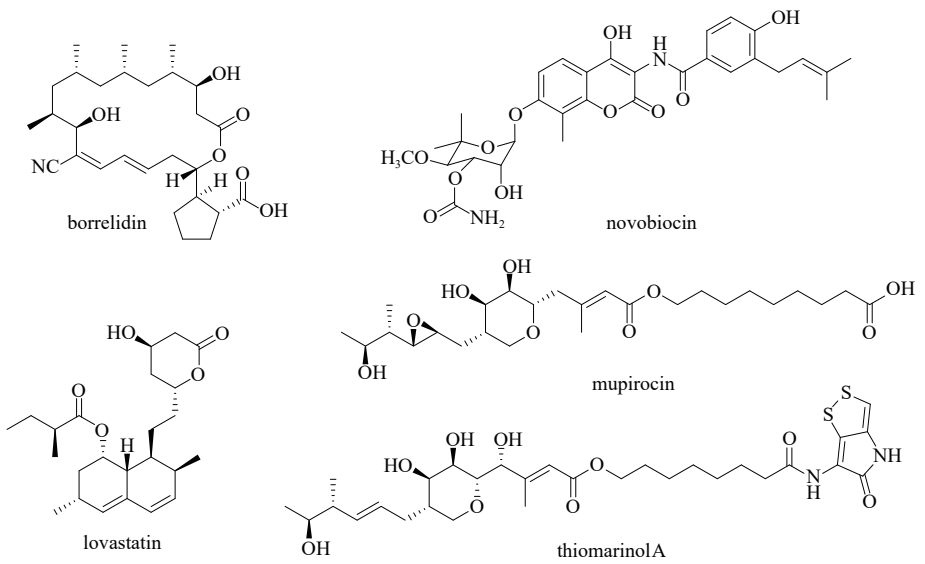


图6 在生物合成基因簇中使用SRE基因编码突变靶点的天然产物

Fig. 6 Natural products that employ a mutated target encoded by SRE genes in biosynthetic gene clusters

一个双拷贝的HMGR，通过实验证实该基因确实具有自抗能力[99-100]。

活性天然产物不仅是人类治疗药物的重要来源，也是许多农业药物的主要来源。支链氨基酸生物合成途径（branched-chain amino acid，BCAA）是植物生长的重要途径，它不存在于动物中，因此是高度特异性除草剂的有效靶点[101]。植物中的BCAA生物合成途径是由三种酶完成的：乙酰乳酸合成酶（acetolactate synthase，ALS）、乙酰羟基异构还原酶（acetohydroxy acid isomeroreductase，KARI）以及二羟基酸脱水酶（dihydroxy-acid dehydratase，DHAD）。DHAD是一种重要且高度保守的植物催化酶，它催化β-脱水反应生成α-酮酸前体，进一步生成异亮氨酸、缬氨酸和亮氨酸，发展DHAD的抑制剂成为制备除草剂的重要工业手段。为了鉴定可能编码DHAD抑制剂的天然产物生物合成基因簇，Tang Yi等[102]利用SRE策略，假定其目标生物合成基因簇中包含一个对抑制剂不敏感的DHAD拷贝，从DHAD出发进行真菌基因组扫描，结合进化树分析等方法从土曲霉（*Aspergillus terreus*）中挖掘到一个与其高度同源的基因*astD*，对其所在的基因簇进行异源表达获得了新颖的天然产物分子，从而发现了一种天然除草剂aspterric acid（图7），并确定了其作用机制。

随着天然产物生物合成基因簇的进化，与其共簇的SRE也会随之而进化，SRE不仅能够为抗生素的耐药性提供新的见解，同时也为抗生素的靶点提供新的切入点。然而，从SRE出发利用现有的知识和信息获得的天然产物，有时并不是我们期待的目标产物[103]，因此，准确地预测SRE还是目前天然产物发现过程中一个极具挑战性的工作。

2.3 基于系统进化进行基因组挖掘

天然产物的结构多样性是生物合成基因簇不断进化的结果。分子系统发育是一种常用的跟踪特定基因序列的进化足迹，并确定其与同源序列的进化关系的技术。以系统发育为导向发现新天然产物的基本思想是根据一个生物合成基因与其各自的生物合成基因簇共同进化，可以作为系统发育标志，代表其整个生物合成基因簇的进化路径，通过进化关系的远近判断天然产物的新颖程度[104]（图8）。

利用系统进化分析挖掘天然产物最成功的案例是芳香聚酮类化合物[105-108]。芳香族聚酮是由Ⅱ型聚酮合酶（PKS）基因簇编码合成的，在Ⅱ型PKS中最小的PKS模块包括酮基合酶α（KS_α）、酮基合酶β（KS_β）和酰基载体蛋白（ACP）[106]。这三个基因参与了芳香族聚酮生物合成过程的第一步，通过催化丙二酰辅酶A（malonyl-CoA）单元

的重复缩合产生不同长度的线性聚酮链（图9）。这些最小的PKS基因可能与它们各自的生物合成基因簇共同进化，因此可以作为系统发育标记。

蒽醌类化合物（anthracyclines）是一类具有抗肿瘤活性的天然产物[109]，其中具有代表性的多柔比星（doxorubicin）已用于临床抗癌化疗超过30年[110]。在系统发育分析中，宏基因组DNA衍生的扩增子序列AZ129与已知的蒽醌类化合物斯

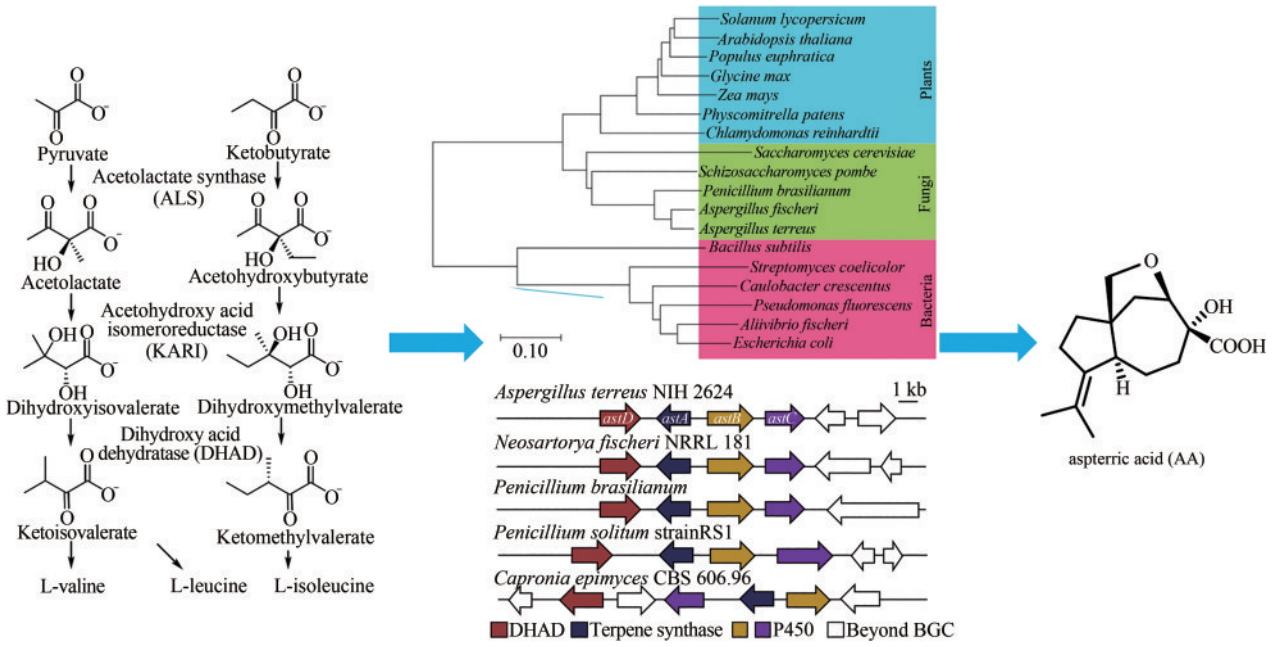


图7 从BCAA生物合成路径中关键的DHAD酶出发挖掘天然除草剂AA[102]

Fig. 7 Genome mining of a natural herbicide aspterric acid (AA) from the critical DHAD enzyme in the BCAA biosynthesis pathways

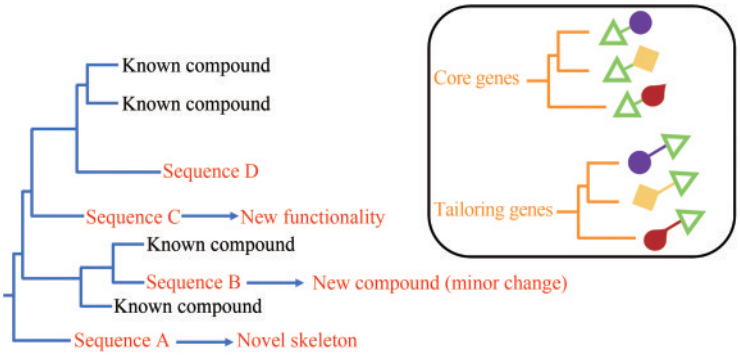


图8 利用标记基因序列建立系统发育树来指导新天然产物的发现[104]

Fig. 8 Phylogenetic tree built with marker gene sequences for guiding the discovery of novel natural products

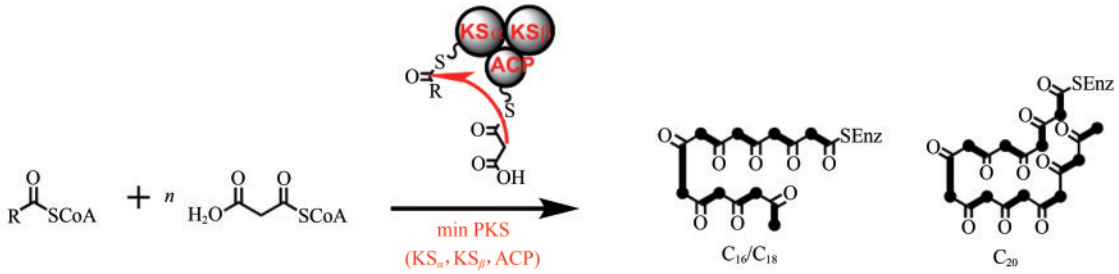


图9 最小化的PKS参与不同长度的线性聚酮链的合成

Fig. 9 minPKS involved in the synthesis of diverse linear polyketide chains

特菲霉素（steffimycin）生物合成基因簇的序列形成一个紧密的分支，Brady等[111]利用AZ129扩增子序列作为探针从宏基因组中重新获得的AZ129基因簇的信息学分析表明，与斯特菲霉素生物合成基因簇相比，还存在一组额外的糖生物合成基因。在白色链霉菌（*Streptomyces albus*）中对AZ129基因簇进行异源表达获得一个全新的天然产物arimetamycin A（图10），在体外肿瘤细胞抗增殖实验中，arimetamycin A表现出比多柔比星更强的活性，并且对多柔比星耐药的癌细胞也表现出中等的抗肿瘤活性[111]。

系统进化分析除了利用上述编码聚酮合成酶这类骨架形成相关的基因作为标签，还可以利用前体供应基因、编码后修饰蛋白的基因、抗性基因等特征基因作为标签，通过进化关系将其与相应的代谢产物关联，指导新结构、新活性天然产物的发现。

3 展 望

大自然从数十亿年前就开始以“自然实验师”的身份进行生物工程实验。为了探究自然界神秘的面纱，人们开发了许多基于基因组与宏基因组的策略来剖析生物进化过程中涉及的途径，并发现了许多新的药物和高效的生物催化剂（酶），同时解析了这些新的反应机制。天然产物及其衍生物一直都是药物先导化合物的重要来源。在天然产物的获取途径中，传统的分离分析方法无法避免重复性、滞后性等问题，这不仅耗时耗力而且无法突破代谢产物“黑箱子”的魔咒。随着基因组测序技术的快速发展，以基因组学为导向的天然产物发现已经成为药物研究领域的重要组成部分。尽管持续更新的基因组数据为天然产物的研究和开发提供了源源不断的资源，然而，如何利用现有的实验条件和技术进行天然产物的挖掘还是一项极具挑战的任务。就微生物领域而言，目前所报道的微生物种群和基因组数据远远超过已知代谢产物的数量，换言之，还有数以万计的“沉默”基因簇等待着开发和利用。以数据为研究中心的方法正在从根本上改变自然科学的许多领域[112]，多组学技术、系统生物学方法与合成生物学理论的联合使用推动着基因组、转录组和代谢组数据的自动化高通量分析，从而更有效地将基因与有机分子连接起来。基于这些理论的结合使用以及网络工具的更新与发展，许多新颖的挖掘技术被开发出来，研究方法已经从传统上基于活性导向天然产物的发现，转向基于核心骨架基因、基于抗性基因以及基于系统进化的基因组挖掘，通过对化学结构、基因组和代谢组学等数据的集成为我们提供了数据的优先排序。这样，基因组的挖掘不仅能发现“终点”药物分子，而且对多个研究领域的发展也起到了非常关键的承接作用。同时，参与次级代谢的酶催化各种各样的反应，这些反应可以在合成生物学中进化和利用。天然产物本身在介导微生物-微生物相互作用、宿主-微生物相互作用以及影响疾病、生长发育等方面都发挥着重要作用。许多合成化学家通过合成结构复杂、活性显著的有机小分子从而开发了许多高效、绿色环保的合成路线，加快了药物合成的步伐。生物学家通过研究生命体内包括转录、调控以及相应的酶学机制，阐明了生命传承过程中许多重要的途径。天然产物研究改革与创新正在创建一种多领域多学科交叉的研究模式，这种方式

图10 利用KS_{β}基因系统发育成功地定向发现新的蒽环类候选药物先导化合物[104-111]

Fig. 10 KS_{β} gene phylogeny used for the targeted discovery of new lead anthracycline compounds

汇聚了多种学习方法、理论基础以及实时更新的网络信息学技术。如今，随着科技的快速发展，人工智能（artificial intelligence）在各个领域都开始崭露头角，在科技时代如何把握技术的更新和运用将成为基因组挖掘研究领域发展的一大挑战。

参 考 文 献

[1] NEWMAN D J, CRAGG G M. Natural products as sources of new drugs from 1981 to 2014[J]. Journal of Natural Products, 2016, 79(3): 629-661.

[2] NEWMAN D J, CRAGG G M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019[J]. Journal of Natural Products, 2020, 83 (3): 770-803.

[3] LERNER C G, HAJDUK P J, WAGNER R, et al. From bacterial genomes to novel antibacterial agents: Discovery, characterization, and antibacterial activity of compounds that bind to HI0065 (YjeE) from *Haemophilus influenzae*[J]. Chemical Biology & Drug Design, 2007, 69 (6): 395-404.

[4] CHALLIS G L. Genome mining for novel natural product discovery[J]. Journal of Medicinal Chemistry, 2008, 51(9): 2618-2628.

[5] ZALLOT R, OBERG N O, GERLT J A. “Democratized” genomic enzymology web tools for functional assignment[J]. Current Opinion in Chemical Biology, 2018, 47: 77-85.

[6] WEBER T. In silico tools for the analysis of antibiotic biosynthetic pathways[J]. International Journal of Medical Microbiology, 2014, 304(3/4): 230-235.

[7] MEDEMA M H, FISCHBACH M A. Computational approaches to natural product discovery[J]. Nature Chemical Biology, 2015, 11(9): 639-648.

[8] UMEMURA M, KOIKE H, MACHIDA M. Motif-independent de novo detection of secondary metabolite gene clusters-toward identification from filamentous fungi[J]. Frontiers in Microbiology, 2015, 6: 371.

[9] WEBER T, KIM H U. The secondary metabolite bioinformatics portal: computational tools to facilitate synthetic biology of secondary metabolite production[J]. Synthetic and Systems Biotechnology, 2016, 1(2): 69-79.

[10] BACHMANN B O, LANEN S G, BALTZ R H. Microbial genome mining for accelerated natural products discovery: Is a renaissance in the making [J]? Journal of Industrial Microbiology & Biotechnology, 2014, 41(2): 175-184.

[11] BODDY C N. Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides[J]. Journal of Industrial Microbiology & Biotechnology, 2014, 41(2): 443-450.

[12] SCHEFFLER R J, COLMER S, TYNAN H, et al. Antimicrobials, drug discovery, and genome mining[J]. Applied Microbiology and Biotechnology, 2013, 97(3): 969-978.

[13] YAEGASHI J, OAKLEY B R, WANG C C C. Recent advances in genome mining of secondary metabolite biosynthetic gene clusters and the development of heterologous expression systems in *Aspergillus nidulans*[J]. Journal of Industrial Microbiology & Biotechnology, 2014, 41(2): 433-442.

[14] VAN SANTEN J A, KAUTSAR S A, MEDEMA M H, et al. Microbial natural product databases: moving forward in the multi-omics era[J]. Natural Product Reports, 2021, 38(1): 264-278.

[15] SOROKINA M, STEINBECK C. Review on natural products databases: where to find data in 2020[J]. Journal of Cheminformatics, 2020, 12(1): 20.

[16] HARBORNE J B. Dictionary of natural products[EB/OL]. 2015, http://dnp. chenmnetbase. com/faces/chemical/ChemicalSearch.xhtml.

[17] BOLTON EVAN E, WANG Y L, THIESSEN P A, et al. PubChem: integrated platform of small molecules and biological activities[J]. Annual Reports in Computational Chemistry, 2010, 4: 217-241.

[18] AFENDI F M, OKADA T, YAMAZAKI M, et al. KNApSAcK family databases: Integrated metabolite-plant species databases for multifaceted plant research[J]. Plant and Cell Physiology, 2012, 53(2): e1.

[19] CABOCHE S, PUPIN M, LECLÈRE V, et al, Norine: a database of nonribosomal peptides[J]. Nucleic Acids Research, 2008, 36(1): D326-D331.

[20] ZIN P P K, WILLIAMS G J, EKINS S. Cheminformatics analysis and modeling with MacrolactoneDB[J]. Scientific Reports, 2020, 10(1): 6284.

[21] ZENG X, ZHANG P, HE W D, et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development[J]. Nucleic Acids Research, 2018, 46(D1): D1217-D1222.

[22] KLEMENTZ D, DÖRING K, LUCAS X, et al. StreptomeDB 2.0—an extended resource of natural products produced by *Streptomycetes*[J]. Nucleic Acids Research, 2016, 44(D1): D509-D514.

[23] VAN SANTEN J A, JACOB G, SINGH A L, et al. The natural products atlas: an open access knowledge base for microbial natural products discovery[J]. ACS Central Science, 2019, 5(11): 1824-1833.

[24] KAUTSAR S A, BLIN K, SHAW S, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function[J]. Nucleic Acids Research, 2020, 48(D1): D454-D458.

[25] WANG M, CARVER J J, PHELAN V V, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking[J]. Nature Biotechnology, 2016, 34(8): 828-837.

[26] BENSON D A, CAVANAUGH M, CLARK K, et al. GenBank[J]. Nucleic Acids Research, 2013, 41(D1): D36-D42.

[27] CONWAY K R, BODDY C N. ClusterMine360: a database of microbial PKS/NRPS biosynthesis[J]. Nucleic Acids Research, 2013, 41(D1): D402-D407.

[28] ICHIKAWA N, SASAGAWA M, YAMAMOTO M, et al. DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters[J]. Nucleic Acids Research, 2013, 41(D1): D408-D414.

[29] CHEN I M A, CHU K, PALANIAPPAN K, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes[J]. Nucleic Acids Research, 2019, 47(D1): D666-D677.

[30] BLIN K, PASCAL ANDREU V, DE LOS SANTOS E L C, et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters[J]. Nucleic Acids Research, 2019, 47(D1): D625-D630.

[31] DIMINIC J, ZUCKO J, RUZIC I T, et al. Databases of the thiotemplate modular systems (CSDB) and their in silico recombinants (R-CSDB)[J]. Journal of Industrial Microbiology & Biotechnology, 2013, 40(6): 653-659.

[32] MEDEMA M H, KOTTMANN R, YILMAZ P, et al. Minimum information about a biosynthetic gene cluster[J]. Nature Chemical Biology, 2015, 11(9): 625-631.

[33] GRIGORIEV I V, NIKITIN R, HARIDAS S, et al. MycoCosm portal: gearing up for 1000 fungal genomes[J]. Nucleic Acids Research, 2014, 42(D1): D699-D704.

[34] BLIN K, MEDEMA M H, KOTTMANN R, et al. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters[J]. Nucleic Acids Research, 2017, 45(D1): D555-D559.

[35] O′LEARY N A, WRIGHT M W, BRISTER J R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation[J]. Nucleic Acids Research, 2016, 44(D1): D733-D745.

[36] The UniProt Consortium. UniProt: The universal protein knowledgebase[J]. Nucleic Acids Research, 2017, 45(D1): D158-D169.

[37] RADIVOJAC P, CLARK W T, ORON T R, et al. A large-scale evaluation of computational protein function prediction[J]. Nature Methods, 2013, 10(3): 221-227.

[38] EL-GEBALI S, MISTRY J, BATEMAN A, et al. The Pfam protein families database in 2019[J]. Nucleic Acids Research, 2019, 47(D1): D427-D432.

[39] BLUM M, CHANG H-Y, CHUGURANSKY S, et al. The InterPro protein families and domains database: 20 years on[J]. Nucleic Acids Research, 2021:49(D1):D344-D354.

[40] FINN R D, COGGILL P, EBERHARDT R Y, et al. The Pfam protein families database: towards a more sustainable future[J]. Nucleic Acids Research, 2016, 44(D1): D279-D285.

[41] BENTLEY S D, CHATER K F, CERDEÑO-TÁRRAGA A M, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2) [J]. Nature, 2002, 417(6885): 141-147.

[42] CAMACHO C, COULOURIS G, AVAGYAN V, et al. BLAST+: architecture and applications[J]. BMC Bioinformatics, 2009, 10: 421.

[43] EDDY S R. Accelerated profile HMM searches[J]. PLoS Computational Biology, 2011, 7(10): e1002195.

[44] BUCHFINK B, XIE C, HUSON D H. Fast and sensitive protein alignment using DIAMOND[J]. Nature Methods, 2015, 12(1): 59-60.

[45] STARCEVIC A, ZUCKO J, SIMUNKOVIC J, et al. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures[J]. Nucleic Acids Research, 2008, 36(21): 6882-6892.

[46] WEBER T, RAUSCH C, LOPEZ P, et al. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters[J]. Journal of Biotechnology, 2009, 140(1/2): 13-17.

[47] LI M H, UNG P M, ZAJKOWSKI J, et al. Automated genome mining for natural products[J]. BMC Bioinformatics, 2009, 10(1): 185.

[48] KHALDI N, SEIFUDDIN F T, TURNER G, et al. SMURF: genomic mapping of fungal secondary metabolite clusters[J]. Fungal Genetics and Biology, 2010, 47(9): 736-741.

[49] MEDEMA M H, BLIN K, CIMERMANCIC P, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences[J]. Nucleic Acids Research, 2011, 39(2): W339-W346.

[50] BLIN K, SHAW S, STEINKE K, et al. antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline[J]. Nucleic Acids Research, 2019, 47(W1): W81-W87.

[51] MICHAEL A, FISCHBACH C T W. Antibiotics for emerging pathogens[J]. Science, 2009, 325: 1089-1093.

[52] STREIT W R, SCHMITZ R A. Metagenomics-the key to the uncultured microbes[J]. Current Opinion in Microbiology, 2004, 7(5): 492-498.

[53] CIMERMANCIC P, MEDEMA M H, CLAESEN J, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters[J]. Cell, 2014, 158(2): 412-421.

[54] CRUZ-MORALES P, KOPP J F, MARTINEZ-GUERRERO C, et al. Phylogenomic analysis of natural products biosynthetic gene clusters allows discovery of arseno-organic metabolites in model streptomycetes[J]. Genome Biology and Evolution,

2016, 8(6):1906-1916.

[55] TAKEDA I, UMEMURA M, KOIKE H, et al. Motif-Independent prediction of a secondary metabolism gene cluster using comparative genomics: Application to sequenced genomes of *Aspergillus* and ten other filamentous fungal species[J]. DNA Research, 2014, 21(4): 447-457.

[56] SANTOS-ABERTURAS J, CHANDRA G, FRATTARUOLO L, et al. Uncovering the unexplored diversity of thioamidated ribosomal peptides in Actinobacteria using the RiPPER genome mining tool[J]. Nucleic Acids Research, 2019, 47(9): 4624-4637.

[57] KLOOSTERMAN A M, SHELTON K E, VAN WEZEL G P, et al. RRE-Finder: A genome-mining tool for class-independent RiPP discovery[J]. mSystems, 2020, 5(5): e00267-00220.

[58] MERWIN N J, MOUSA W K, DEJONG C A, et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(1): 371-380.

[59] TIETZ J I, SCHWALEN C J, PATEL P S, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape[J]. Nature Chemical Biology, 2017, 13(5): 470-478.

[60] HASTINGS J, DE MATOS P, DEKKER A, et al. The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013[J]. Nucleic Acids Research, 2013, 41(D1): D456-D463.

[61] BENTO A P, GAULTON A, HERSEY A, et al. The ChEMBL bioactivity database: an update[J]. Nucleic Acids Research, 2014, 42(D1): D1083-D1090.

[62] PENCE H E, WILLIAMS A. ChemSpider: an online chemical information resource[J]. Journal of Chemical Education, 2010, 87(11): 1123-1124.

[63] KAUTSAR S A, BLIN K, SHAW S, et al. BiG-FAM: the biosynthetic gene cluster families database[J]. Nucleic Acids Research, 2020, 49(D1): D490-D497.

[64] WEN L. Progress of biosynthesis of enediyne antitumor antibiotics[J]. World Sci-Tech R & D, 2005, 27: 3.

[65] ZAZOPOULOS E, HUANG K, STAFFA A, et al. A genomics-guided approach for discovering and expressing cryptic metabolic pathways[J]. Nature Biotechnology, 2003, 21(2): 187-190.

[66] ADHIKARI A, TEIJARO C N, TOWNSEND C A, et al. 1.12-biosynthesis of enediyne natural products[M]//LIU H-W, BEGLEY T P. Comprehensive natural products iii. Oxford: Elsevier, 2020: 365-414.

[67] SHEN B, HINDRA, YAN X, et al. Enediynes: exploration of microbial genomics to discover new anticancer drug leads[J]. Bioorganic & Medicinal Chemistry Letters, 2015, 25(1): 9-15.

[68] YAN X, GE H, HUANG T, et al. Strain prioritization and genome mining for enediyne natural products[J]. mBio, 2016, 7(6): e02104-02116.

[69] ORTEGA M A, HAO Y, ZHANG Q, et al. Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB[J]. Nature, 2015, 517(7535): 509-512.

[70] HUDSON G A, ZHANG Z G, TIETZ J I, et al. *In vitro* biosynthesis of the core scaffold of the thiopeptide thiomuracin[J]. Journal of the American Chemical Society, 2015, 137(51): 16012-16015.

[71] TING C P, FUNK M A, HALABY S L, et al. Use of a scaffold peptide in the biosynthesis of amino acid-derived natural products[J]. Science, 2019, 365(6450): 280-284.

[72] HAMADA T, MATSUNAGA S, FUJIWARA M, et al. Solution structure of polytheonamide B, a highly cytotoxic nonribosomal polypeptide from marine sponge[J]. Journal of the American Chemical Society, 2010, 132(37): 12941-12945.

[73] MORINAKA B I, VAGSTAD A L, HELF M J, et al. Radical *S*-adenosyl methionine epimerases: Regioselective introduction of diverse D-amino acid patterns into peptide natural products[J]. Angewandte Chemie International Edition, 2014, 53(32): 8503-8507.

[74] FREEMAN M F, GURGUI C, HELF M J, et al. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides[J]. Science, 2012, 338(6105): 387-390.

[75] BHUSHAN A, EGLI P J, PETERS E E, et al. Genome mining- and synthetic biology-enabled production of hypermodified peptides[J]. Nature Chemistry, 2019, 11(10): 931-939.

[76] BRODERICK J B, DUFFUS B R, DUSCHENE K S, et al. Radical *S*-adenosylmethionine enzymes[J]. Chemical Reviews, 2014, 114(8): 4229-4317.

[77] BENJDIA A, GUILLOT A, LEFRANC B, et al. Thioether bond formation by SPASM domain radical SAM enzymes: C_α H-atom abstraction in subtilosin A biosynthesis[J]. Chemical Communications, 2016, 52(37): 6249-6252.

[78] HUDSON G A, BURKHART B J, DICAPRIO A J, et al. Bioinformatic mapping of radical *S*-adenosylmethionine-dependent ribosomally synthesized and post-translationally modified peptides identifies new C_α, C_β, and C_γ-linked thioether-containing peptides[J]. Journal of the American Chemical Society, 2019, 141(20): 8228-8238.

[79] SCHRAMMA K R, BUSHIN L B, SEYEDSAYAMDOST M R. Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink[J]. Nature Chemistry, 2015, 7(5): 431-437.

[80] GARDAN R, BESSET C, GUILLOT A, et al. The Oligopeptide transport system is essential for the development of natural competence in *Streptococcus thermophilus* strain LMD-9[J]. Journal of Bacteriology, 2009, 191(14): 4647-4655.

[81] BUSHIN L B, CLARK K A, PELCZER I, et al. Charting an

unexplored streptococcal biosynthetic landscape reveals a unique peptide cyclization motif[J]. Journal of the American Chemical Society, 2018, 140(50): 17674-17684.

[82] CARUSO A, MARTINIE R J, BUSHIN L B, et al. Macrocyclization *via* an arginine-tyrosine crosslink broadens the reaction scope of radical *S*-adenosylmethionine enzymes[J]. Journal of the American Chemical Society, 2019, 141(42): 16610-16614.

[83] CLARK K A, BUSHIN L B, SEYEDSAYAMDOST M R. Aliphatic ether bond formation expands the scope of radical SAM enzymes in natural product biosynthesis[J]. Journal of the American Chemical Society, 2019, 141(27): 10610-10615.

[84] CARUSO A, BUSHIN L B, CLARK K A, et al. Radical approach to enzymatic β-thioether bond formation[J]. Journal of the American Chemical Society, 2019, 141(2): 990-997.

[85] BUSHIN L B, COVINGTON B C, RUED B E, et al. Discovery and biosynthesis of streptosactin, a sactipeptide with an alternative topology encoded by commensal bacteria in the human microbiome[J]. Journal of the American Chemical Society, 2020, 142(38): 16265-16275.

[86] LAUTRU S, DEETH R J, BAILEY L M, et al. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining[J]. Nature Chemical Biology, 2005, 1(5): 265-269.

[87] YAN Y, LIU N, TANG Y. Recent developments in self-resistance gene directed natural product discovery[J]. Natural Product Reports, 2020, 37(7): 879-892.

[88] GALM U, HAGER M H, VAN LANEN S G, et al. Antitumor antibiotics: bleomycin, enediynes, and mitomycin[J]. Chemical Reviews, 2005, 105(2): 739-758.

[89] WEISBLUM B. Erythromycin resistance by ribosome modification[J]. Antimicrobial Agents and Chemotherapy, 1995, 39(3): 577-585.

[90] O'NEILL E C, SCHORN M, LARSON C B, et al. Targeted antibiotic discovery through biosynthesis-associated resistance determinants: Target directed genome mining[J]. Critical Reviews in Microbiology, 2019, 45(3): 255-277.

[91] ALMABRUK K H, DINH L K, PHILMUS B. Self-resistance of natural product producers: past, present, and future focusing on self-resistant protein variants[J]. ACS Chemical Biology, 2018, 13(6): 1426-1437.

[92] MAXWELL A. The interaction between coumarin drugs and DNA gyrase[J]. Molecular Microbiology, 1993, 9(4): 681-686.

[93] THIARA A S, CUNDLIFFE E. Interplay of novobiocin-resistant and -sensitive DNA gyrase activities in self-protection of the novobiocin producer, *Streptomyces sphaeroides*[J]. Gene, 1989, 81(1): 65-72.

[94] STEFFENSKY M, MÜHLENWEG A, WANG Z-X, et al. Identification of the novobiocin biosynthetic gene cluster of *Streptomyces spheroides* NCIB 11891[J]. Antimicrobial Agents and Chemotherapy, 2000, 44(5): 1214-1222.

[95] EL-SAYED A K, HOTHERSALL J, COOPER S M, et al. Characterization of the mupirocin biosynthesis gene cluster from *Pseudomonas fluorescens* NCIMB 10586[J]. Chemistry & Biology, 2003, 10(5): 419-430.

[96] FUKUDA D, HAINES A S, SONG Z, et al. A natural plasmid uniquely encodes two biosynthetic pathways creating a potent anti-MRSA antibiotic[J]. PLoS One, 2011, 6(3): e18031.

[97] OLANO C, WILKINSON B, SÁNCHEZ C, et al. Biosynthesis of the angiogenesis inhibitor Borrelidin by *Streptomyces parvulus* Tü4055: Cluster Analysis and Assignment of Functions [J]. Chemistry & Biology, 2004, 11(1): 87-97.

[98] STANCU C, SIMA A. Statins: mechanism of action and effects[J]. Journal of Cellular and Molecular Medicine, 2001, 5(4): 378-387.

[99] HUTCHINSON C R, KENNEDY J, PARK C, et al. Aspects of the biosynthesis of non-aromatic fungal polyketides by iterative polyketide synthases[J]. Antonie van Leeuwenhoek, 2000, 78(3/4): 287-295.

[100] CHAMILOS G, LEWIS R E, KONTOYIANNIS D P. Lovastatin has significant activity against zygomycetes and interacts synergistically with voriconazole[J]. Antimicrobial Agents and Chemotherapy, 2006, 50(1): 96-103.

[101] AMORIM FRANCO T M, BLANCHARD J S. Bacterial branched-chain amino acid biosynthesis: structures, mechanisms, and drugability[J]. Biochemistry, 2017, 56(44): 5849-5865.

[102] YAN Y, LIU Q, ZANG X, et al. Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action[J]. Nature, 2018, 559(7714): 415-418.

[103] TANG M C, ZOU Y, YEE D, et al. Identification of the pyranonigrin A biosynthetic gene cluster by genome mining in *Penicillium thymicola* IBT 5891[J]. AIChE Journal, 2018, 64(12): 4182-4186.

[104] KANG H-S. Phylogeny-guided (meta) genome mining approach for the targeted discovery of new microbial natural products[J]. Journal of Industrial Microbiology & Biotechnology, 2017, 44(2): 285-293.

[105] FENG Z, KALLIFIDAS D, BRADY S F. Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites[J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(31): 12629-12634.

[106] HERTWECK C. The biosynthetic logic of polyketide diversity[J]. Angewandte Chemie International Edition, 2009, 48(26): 4688-4716.

[107] HILLENMEYER M E, VANDOVA G A, BERLEW E E, et al. Evolution of chemical diversity by coordinated gene swaps in type II polyketide gene clusters[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(45): 13952-13957.

[108] KANG H-S, BRADY S F. Mining soil metagenomes to better

understand the evolution of natural product structural diversity: Pentangular polyphenols as a case study[J]. Journal of the American Chemical Society, 2014, 136(52): 18111-18119.

[109] MINOTTI G, MENNA P, SALVATORELLI E, et al. Anthracyclines: Molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity[J]. Pharmacological Reviews, 2004, 56(2): 185.

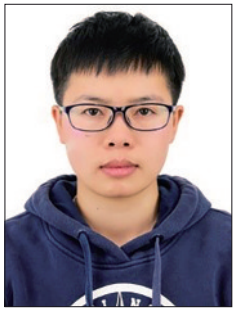
[110] TACAR O, SRIAMORNSAK P, DASS C R. Doxorubicin: An update on anticancer molecular action, toxicity and novel drug delivery systems[J]. Journal of Pharmacy and Pharmacology, 2013, 65(2): 157-170.

[111] KANG H S, BRADY S F. Arimetamycin A: improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes[J]. Angewandte Chemie International Edition, 2013, 52(42): 11063-11067.

[112] KANEHISA M, FURUMICHI M, TANABE M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs[J]. Nucleic Acids Research, 2017, 45(D1): D353-D361.

通讯作者：刘文（1971—），男，研究员，博士生导师。研究方向为复杂天然产物的生物合成（遗传学、生物化学和化学），以产量提高和结构多样性为目的组合生物合成，以基因组扫描为手段的新型天然产物发现。

E-mail：wliu@mail.sioc.ac.cn



第一作者：杨谦（1994—）女，博士，博士后。研究方向为天然产物化学及以基因组扫描为手段的新型天然产物发现。

E-mail：yangqian117@sioc.ac.cn