特约评述

DOI: 10.12211/2096-8280.2022-075

微生物组生物合成基因簇发掘方法及应用前景

赖奇龙，姚帅，查毓国，白虹，宁康
（华中科技大学生命科学与技术学院，分子生物物理教育部重点实验室，生物信息与分子成像湖北省重点实验室，人工智能生物学研究中心，生物信息与系统生物学系，湖北　武汉　430074）

摘要：生物合成基因簇（biosynthetic gene cluster， BGC）是一类非常重要的基因集合（gene set）类型。BGC普遍存在于各类生物基因组中，并且发挥着重要的代谢和调控作用。从线性结构上来说，一个BGC中的基因通常在基因组中处于相邻的位置；从基因功能上来说，一个BGC中的基因通常共同负责一类通路，生成特定的化合物小分子。因此，BGC作为极具潜力的元件来源，在合成生物学研究中极为重要。然而从序列模式上来说，一个BGC中的基因数量众多且序列差异度大，很难通过序列同源性发掘新类型的BGC。因此，建立生物合成基因簇的智能发掘策略，系统性地发掘BGC并进行验证和转化研究，不论在理论方面还是实际应用方面，都具有非常重要的价值。本文主要基于微生物组大数据，较全面地介绍了BGC挖掘的意义和瓶颈问题，系统性地总结了当前BGC发掘中的数据资源和挖掘方法，尤其是人工智能方法，指出了干湿结合方法对于验证新发掘BGC的重要价值，同时展示了新发掘BGC的多样性和广泛应用领域。最后，展望了结合现有BGC挖掘方法和合成生物学转化，将如何在广度和宽度方面扩展目前的合成生物学研究。

关键词：生物合成基因簇；人工智能；合成生物学；微生物组

中图分类号：Q93　　**文献标志码：**A

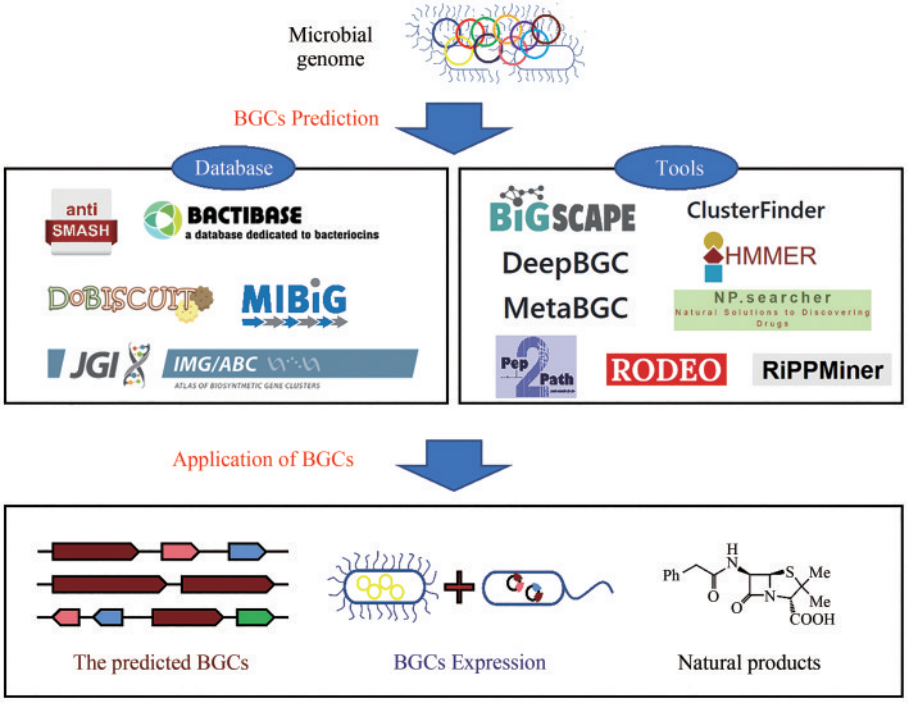
Microbiome-based biosynthetic gene cluster data mining techniques and application potentials

LAI Qilong，YAO Shuai，ZHA Yuguo，BAI Hong，NING Kang
（*Key Laboratory of Molecular Biophysics of the Ministry of Education*， *Hubei Key Laboratory of Bioinformatics and Molecular-imaging*， *Center of Artificial Intelligence Biology*， *Department of Bioinformatics and Systems Biology*， *College of Life Science and Technology*， *Huazhong University of Science and Technology*， *Wuhan 430074*， *Hubei*， *China*）

Abstract: Biosynthetic gene cluster (BGC) is an important type of gene set, which is commonly found in the genomes of various organisms, and plays important metabolic and regulatory roles. In terms of linear gene structure, the set of genes in a BGC is usually located in close proximity to each other in the genome, but for functions, genes in a

收稿日期：2022-12-26 修回日期：2022-03-10
基金项目：国家自然科学基金（32071465，31871334，31671374）；国家重点研发计划（2021YFA0910500）
引用本文：赖奇龙，姚帅，查毓国，白虹，宁康. 微生物组生物合成基因簇发掘方法及应用前景[J]. 合成生物学，2023，4(3)：611-627
Citation：LAI Qilong，YAO Shuai，ZHA Yuguo，BAI Hong，NING Kang. Microbiome-based biosynthetic gene cluster data mining techniques and application potentials[J]. Synthetic Biology Journal，2023，4(3)：611-627

BGC usually work synergistically and are responsible for a class of pathways that generate specific small molecules. Therefore, BGCs are vital in synthetic biology research as a highly promising source for elements. However, current BGC databases and analytical platforms are limited by the number and types of experimentally validated BGCs, as well as by the preliminary BGC data mining techniques. The establishment of data-driven systematic discovery of BGCs and their validation, as well as translational studies, are of great value in both fundamental research and practical applications. This article focuses on mining BGCs from big data with microbiome for synthetic biology research. We start with discussing the definition and significance of BGC mining, and summarize current data resources and methods for BGC mining: including MIBiG, antiSMASH and IMG-ABC for artificial intelligence (AI) enabled web services to accelerate BGC mining. Then, we compile a walk-through on how a typical BGC data mining could be conducted, with the history of BGC mining methods highlighted, which underlines the route build-up from traditional machine learning to deep learning. We also diagnose bottlenecks in BGC mining, and propose possible solutions. Furthermore, according to several BGC mining and validation experiments, we demonstrate the profound diversity and breadth of application scenarios with BGC discovery, as well as the importance of combining dry and wet lab experiments for validating newly discovered BGCs. Finally, we envision that the combination of advanced BGC mining methods and synthetic biology could broaden and deepen current synthetic biology research.



Keywords: biosynthetic gene cluster; artificial intelligence; synthetic biology; microbiome

1 生物合成基因簇：序列与功能

天然产物（natural product， NP）是指生物体内的组成成分或其代谢产物，具有广泛的应用价值[1]，其中源自微生物的次级代谢产物，在生物医学、工业和农业应用中具有重要意义[2]。然而，由于大量环境微生物无法培养[3]，因此挖掘生物合成基因簇（biosynthetic gene cluster， BGC）以检验并生产新型NP当前仍十分困难[4]。在过去的数十年里，随着高通量测序技术和生物大数据处

理工具的快速发展，直接从宏基因组（metagenome）中探索 BGC 的策略已经越来越成熟[5]，这极大地加快了从不可培养微生物（包括极端微生物和稀有微生物等）中发掘新型 BGC 的进度[6]。

生物合成基因簇是一类非常重要的基因集合（gene set）类型。一个 BGC 通常包含数个到上百个功能基因，共同产生一个或者若干个小分子代谢物[7]。例如，合成青霉素的一系列基因，就共同组成了一个 BGC[8]。从现有实验验证过的 BGC 来看，BGC 在序列上和功能上均有鲜明的特征：

从序列上来说，一般情况下，一个 BGC 所囊括的基因，即参与代谢途径中生物合成酶的基因在染色体上成簇排列[9]。例如，青霉素的合成由三个基因控制，分别是 *pcbAB*、*pcbC* 和 *penDE*，这三个基因位于同一条染色体上[10]［图 1(a)］。

从功能上来说，一个 BGC 所囊括的基因，通常共同产生一个或者若干个小分子化合物[11]［图 1(b)］。次生代谢产物（secondary metabolites，SM）是 BGC 合成的主要产物[12]，大部分具有生物活性，通常是低分子量的化合物，在生长和发育的特定阶段产生，这类分子最知名的临床应用包括抗生素（如青霉素）、免疫抑制剂（如环孢菌素）等[13]。又例如，翻译后修饰核糖体多肽（ribosomally synthesized post-translationally modified peptide，RiPP），是由核糖体合成，经由翻译后修饰得到的一大类天然产物，具有广泛的结构和生物活性多样性[14]。由于其化学结构比其他天然产物更具基因组学数据上的可预测性，因此可以通过识别编码 RiPP 的 BGC，在宏基因组中发现新型的 RiPP[15]。

现有数据库中的 BGC 通常是通过湿实验确定的。例如，MIBiG 数据库[16]详细记录了来自于上千个微生物物种的上千个经实验验证的非冗余 BGC。实验验证的工作流程包括新型天然产物的发现和生物合成，这种手段极大地促进了丰富但尚未开发微生物 BGC 的挖掘[17]。在来自世界各地科学家的共同贡献下，MIBiG 数据库于近期又有

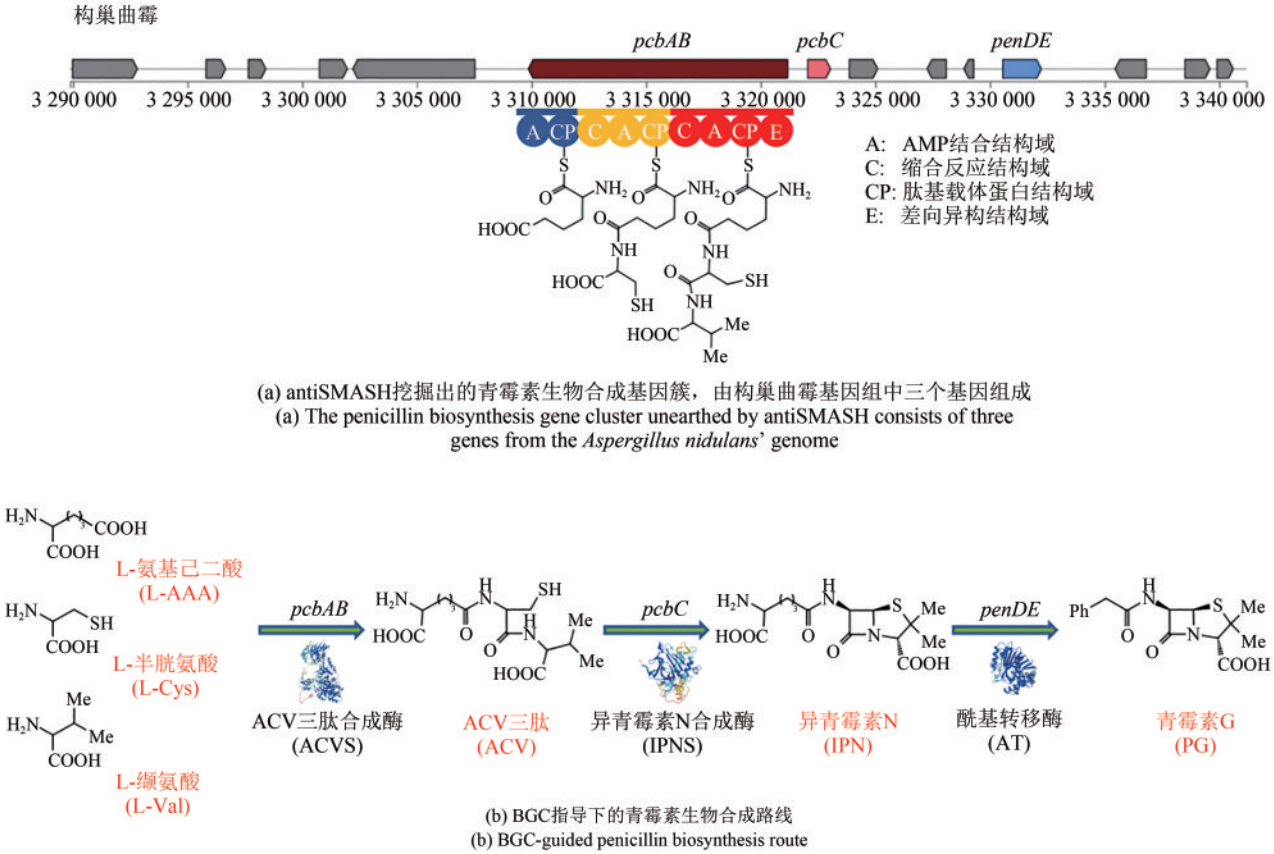


图 1　BGC 在序列和功能上的特征示意图（以青霉素的生物合成为例）

Fig. 1　Schematic diagram for sequences and functions of BGC (with penicillin biosynthesis as an example)

更新，包括2019年新增的851个条目[18]，以及2022年对现有条目的重新注释与661个新条目的大规模验证[19]，目前该数据库收录了2502条已验证的BGC信息。

然而，基于湿实验确定BGC非常复杂且费时，因此一些BGC数据库和计算机比对方法应运而生，如基于局部比对算法的搜索工具（basic local alignment search tool， BLAST）[20]与隐马尔可夫模型（hidden Markov model， HMM）[21]。通过数据库的搜索，能够较为便捷地在基因组中发掘跟已知BGC同源的BGC。例如，antiSMASH数据库[22]（https：//antismash.secondarymetabolites.org/）中包含了所有NCBIGenBank数据库上公布（截止至2022年11月17日）的可用细菌基因组信息（25 802生物物种的82 855条信息）。antiSMASH数据库为研究者提供了一个使用方便、注释了生物合成基因簇的最新集合，以及配套的进行生物合成基因簇搜索分析的方法。然而，针对已知BGC的远源BGC，当前基于数据库的同源搜索尚不能完全胜任。近年来，基于机器学习和深度学习的方法以预测核糖体合成和翻译后修饰肽（RiPP）为重点的方法迅猛增加[23]。下文将通过详细的实例阐明机器学习方法的特点以及其在BGC挖掘中的应用，如metaBGC[24]和DeepBGC[25]等。

2 基于微生物组的生物合成基因簇挖掘与转化研究

许多微生物的次级代谢产物具有抗真菌、抗细菌、抗肿瘤等生物活性，是微生物药物开发和新药创制的重要来源[26]。目前，放线菌和黏细菌等是细菌次级代谢调控和天然产物发掘的重要研究对象[27]。但是，目前对于细菌能合成多少种次级代谢产物、不同类群的细菌在合成次级代谢产物能力方面的差异以及次级代谢产物生物合成基因簇（以下简称次级代谢基因簇）如何进化等问题，尚存在很多未知规律和模式，仍有待研究[28]。

当前，由于BGC转化应用具有广泛的应用价值，重要的BGC通常通过干湿实验共同确定[29]。例如，2022年武汉大学药学院刘天罡课题组[30]开发了“基因簇功能元件理性可控重组”策略，实现了萜类沉默基因簇的批量挖掘及高效合成。这一工作展示了以“基因簇功能元件理性可控重组”策略为指导，从微生物基因组数据出发，进行新化合物挖掘、筛选并实现目标产物高效合成的巨大优势。该项工作详细介绍了从基因组挖掘到萜类化合物生物合成与鉴定的全套流程，为利用人工智能方法（antiSMASH）加速发现微生物组中新型天然产物提供了良好的示范。

目前，有相当多的基于微生物组BGC挖掘和转化的研究项目已经或正在开展[24，31-44]。例如，针对海洋微生物群落进行挖掘，发现了一类全新的海洋细菌（*Candidatus eudoremicrobiaceae*），并预测了近4万种潜在的生物合成基因簇[32]。又比如，针对肠道微生物群落的挖掘，发现了肠道菌群能产生大量不同结构和生物活性的次生代谢产物，与肠道菌分泌的抑菌肽小菌素类似，这些次生代谢产物在药物研发与临床上有很广泛的应用前景[41]。再比如，针对土壤微生物群落进行挖掘，通过对生长在抑病土壤中的甜菜幼苗根进行宏基因组测序分析，区分出哪些BGC在感染过程中表达增加，并通过位点定向诱变分析检验其重要程度，发现抑病土壤中的植物益生菌通过增强真菌细胞壁降解相关酶的活性，为植物提供额外保护[38]。此外，针对特定的微生物，BGC挖掘结果揭示了放线菌基因组具有巨大的天然产物合成潜力[36]，其生产的抗生素在临床中应用前景光明。

3 BGC的分析和比对

BGC的分析和比对，主要是建立在BGC数据库基础之上。大多数BGC数据库提供网页端入口，提交目标序列之后，服务器会根据同源性比对或隐马尔可夫预测等方法展示出最为相似的现有数据，通过解读结果的注释信息即可辅助BGC的分析与预测（图2）。

在BGC数据资源方面，当前服务于不同目的的BGC数据库都有较为广泛和频繁的访问和应用（表1）。

例如，BiG-FAM数据库[47]从公开来源获取了1 225 071个BGC，并使用BiG-SLiCE[53]软件将其

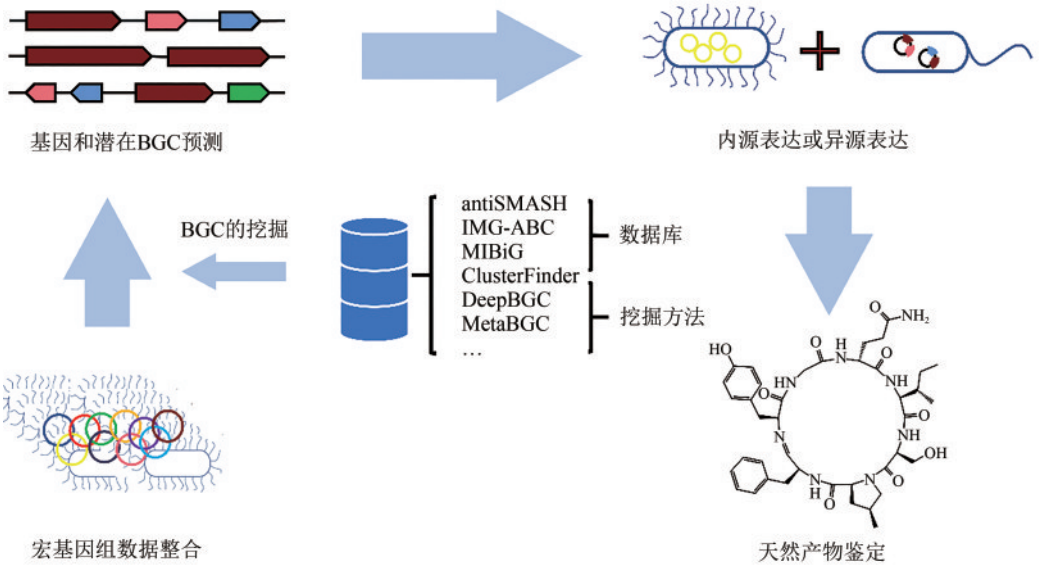


图 2　BGC 挖掘的整体过程

（该过程包括：宏基因组数据的整合，基因和潜在 BGC 的预测，内源表达或异源表达、天然产物的鉴定等。本图中选用的案例是诺糖环肽 A2，是从地衣 *Nostoc* 属 ATCC53789 中提取分离的天然产物，可作为 20S 蛋白酶体的抑制剂，具有抗癌活性 [45]）

Fig. 2　Overall process for BGC mining

(This process includes the integration of metagenomic data, prediction of genes and potential BGC, endogenous or heterologous expression, identification of natural products, *etc*. The case chosen in this figure is Nostocyclopeptide A2, which is extracted from *Nostoc* sp. ATCC53789 isolated from lichen. It can be used as an inhibitor of 20S proteasome and exhibits anticancer activity[45].)

表 1　代表性 BGC 数据库介绍

Table 1　Summary for representative BGC databases

数据库名称	特色	网址	参考文献
antiSMASH	有关次生代谢物 BGC 的综合资源，集成各种分析工具	https://antismash.secondarymetabolites.org/	[22]
Bactibase	主要包括细菌及其产生的抗菌肽、细菌素等	http://bactibase.pfba-lab-tun.org/	[46]
BiG-FAM	将同源 BGCs 分组到基因簇家族	https://bigfam.bioinformatics.nl/	[47]
ClusterMine360	第一个已知产物的 BGC 数据库	http://www.clustermine360.ca/	[48]
CSDB(ClustScan Database)	主要内容为 PKS、NRPS 的 BGC	http://csdb.bioserv.pbf.hr/csdb/ClustScanWeb.html	[49]
DoBISCUIT	提供由文献给出的 PKS 和 NRPS 的 BGC	http://www.bio.nite.go.jp/pks/	[50]
IMG-ABC	最大的公开预测的 BGC 数据库	https://img.jgi.doe.gov/abc-public	[51]
MiBiG	存储 BGC 的最小信息	https://mibig.secondarymetabolites.org/	[19]
OrphanPKS	由软件自动提取的多模块 PKS 序列目录	http://sequence.stanford.edu/OrphanPKS/	[52]

聚类为 29 955 个基因簇家族模型。又例如，IMG-ABC 数据库 [51] 包含了 411 412 个预测 BGC，其中 1332 个 BGC 已得到实验验证，14 985 个 BGC 是从高质量的宏基因组数据中预测得到（截止到 2022 年 12 月）。特定类型的 BGC 数据库如 Bactibase [46]，则覆盖了由 206 种革兰氏阳性菌和 19 种革兰氏阴性菌产生的 230 种抗菌肽或细菌素的 BGC 信息。

在 BGC 比对方法方面，主要包括序列比对和特征比对，多数 BGC 数据库通常都提供了这两种方法进行比对（图 3）。

例如，antiSMASH 数据库 [55] 中提供基于 BLAST 的 ClusterBlast 工具，能将目的基因簇与数据库中的其他基因簇进行序列比对，展示相似性得分最高的多个结果，辅助判断 BGC 的功能与进化上的联系。antiSMASH 数据库还提供了 HMMer3 工具 [56]，可以由基于群落画像（community profile）的隐马尔可夫模型（profile hidden Markov model，pHMM）[57] 刻画特征，与目的序列进行特征比对，检测目的序列中多个特定蛋白质结构域存在的可能性，从而判断出 BGC。

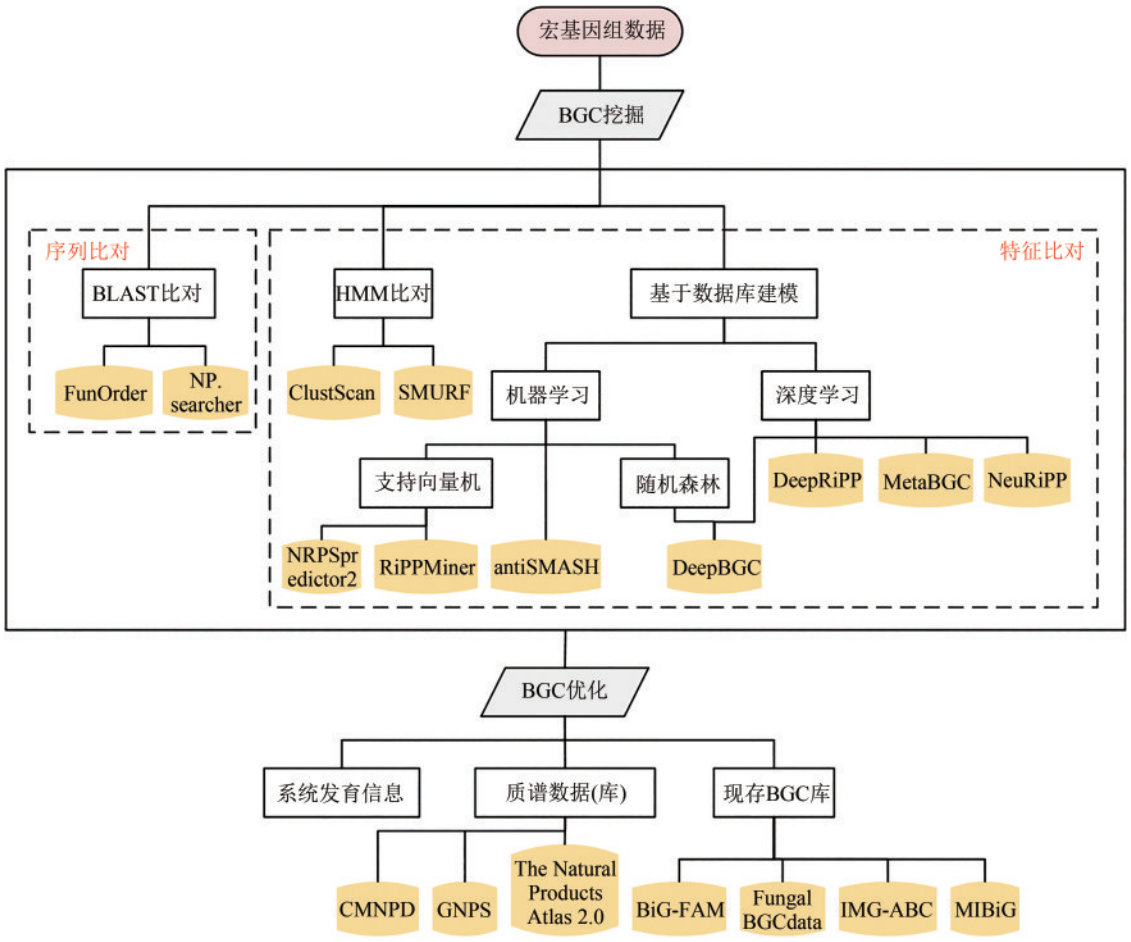


图3 BGC挖掘的一般分析流程及相关方法

[从宏基因组数据中挖掘BGC，主要包括：BGC的挖掘方法（序列比对、特征比对等）和BGC的优化方法（数据库搜索、进化分析等）。其中BGC的挖掘方法主要有序列比对和特征比对两大类：序列比对主要是BLAST等方法，特征比对既包括隐马尔科夫模型（HMM）比对等传统方法，也包括基于数据模型的深度学习等方法。其中BGC的优化方法主要有数据库搜索、进化分析等：数据库搜索包括BGC序列数据库的搜索，以及BGC相关小分子质谱数据库的搜索，而进化分析的主要目标是分析BGC的演化和变异模式[54]]

Fig. 3 Overall flow for BGC analysis and mining

[It mainly includes: BGC mining methods (sequence alignment, feature characterization, etc.) and BGC optimization methods (database searching, evolutionary analysis, *etc.*). Among them, the mining methods of BGC mainly include sequence alignment and feature characterization. Sequence alignment mainly uses BLAST and other methods, while feature characterization employs both traditional methods such as hidden Markov model (HMM) alignment and deep learning based on data model. The optimization methods of BGC mainly include database searching, evolutionary analysis, etc. Database searching includes the searching of BGC sequence database and BGC related small molecule mass spectrometry database, and the main purpose of evolutionary analysis is to analyze the evolution and variation patterns of BGC[54].]

次生代谢产物是BGC合成的主要产物，因此构建序列比对和特征比对方法，将次生代谢产物与其对应BGC联系起来也是计算分析中非常重要的一部分内容（图4）。

当在某个物种中发现了未知的次生代谢产物时，可以先找到与其结构相似且基因簇已被确定的化合物，再根据已知的基因簇通过构建序列比对或特征比对等同源搜索的方式，确定出产生该未知次生代谢产物的候选基因簇。而从BGC确定其次生代谢产物的验证过程，则要利用如异源表达、激活沉默基因等基因工程的手段合成一系列次生代谢产物，其验证方法本文暂不拓展。

4 BGC挖掘的人工智能方法

BGC本质上是基因组编码的遗传信息集合，主要是通过序列数据的分析方法进行分析。因此

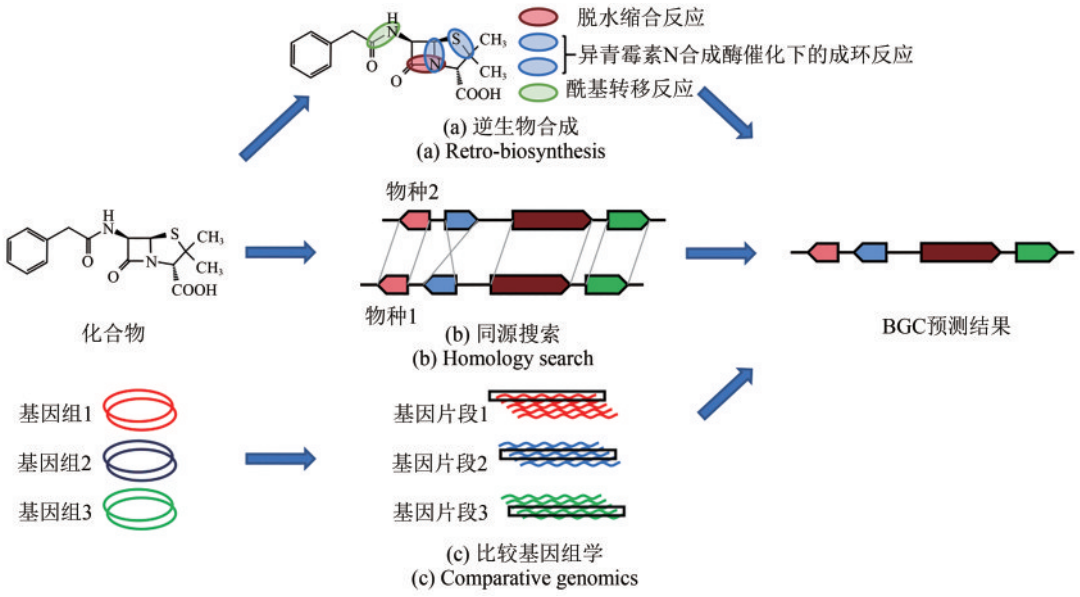


图 4　建立 BGC 和次生代谢产物关联性的分析方法 [58]

（a）逆生物合成：从已知化合物开始，预测生产该化合物所需的活性酶（主干酶和裁剪酶），并从这些预测中找到与基因组中需求匹配的假定簇。本图中选用的案例为青霉素 G [59]。（b）同源搜索：从物种 1 产生的已知化合物和物种 2 产生的相同或相似的化合物开始，使用来自物种 2 的已知基因集群在物种 1 的基因组中搜索相似的基因集群，从而确定感兴趣的基因集群。（c）比较基因组学：从一组生物开始，其中一些生物产生目标化合物，而另一些生物则不产生，有可能在生产中识别同源基因簇，并在非生产中没有同源基因的基础上进行筛选，从而识别候选基因簇

Fig. 4　Analytical methods for establishing correlation between BGC and the production of secondary metabolites[58]

(a) Retro-biosynthesis: starting with a known compound but no related gene clusters identified, it is possible for predicting enzyme(s) to catalyze the synthesis of such a compound (backbone and tailoring enzymes), and with these predictions putative gene clusters matching the requirements can be found in the genome. The selected case in this figure is penicillin G[59]. (b) Homology searching: starting with a known compound produced by organism 1 and the same or similar compound produced by organism 2 with gene cluster identified, it is possible to use the known gene cluster from organism 2 to search for a similar gene cluster in the genome of organism 1, and thereby identify the gene cluster of interest. (c) Comparative genomics: starting with a group of organisms, some of which produce compounds of interest and some of which do not, it is possible to identify homologous gene clusters in the species that produce them and to screen on the basis of the absence of homologous genes in the species that does not produce them, thereby identifying candidate gene clusters.

序列分析的人工智能方法，在很大程度上涵盖了挖掘 BGC 的人工智能方法，其中成熟的方法对 BGC 的人工智能挖掘具有较高的借鉴与参考价值。

4.1　序列分析的人工智能方法

随着生物大数据规模的不断提高，针对生物大数据分析的人工智能（artificial intelligence，AI）方法层出不穷 [60]。目前，AI 技术在生物医药领域应用主要包括药物研发、医学影像、辅助诊疗和基因分析四个子领域。其中，国外借助先进的药品研发技术和人工智能技术起步更早，以 AI 药物研发为主 [61]；我国则借助海量大数据的优势，以 AI 医学影像为主 [62]。大数据可以减少临床研究中的试错成本、大大加快临床实验的成功，也可以集成患者的信息，生成无数生物数据模型，帮助人类理解生命奥秘，实现疾病的精准判断与精准治疗。人工智能可能用人类无法实现的方式整合或解开复杂的基因组数据或是帮助研究者寻找纷繁复杂实验数据中的规律、理解疾病在组学层面的时空动态模式，将为新药研发、临床研究、治疗模式等各方面带来翻天覆地的变革 [63]。

序列分析的人工智能方法 [64]，是人工智能在生物序列分析特定场景下的方法，包括 PICS [65]、DeepCell [66] 等图像识别方法，Enformer [67]、DeepLinc[68]等基因表达分析方法，以及 AlphaFold2[69] 等结构功能预测方法。基因二代测序技术产生了大量的测序数据，AI 在基因大数据的分析上亦表现出良好的和不断扩展的应用趋势（图 5），即在分子层面的基因组学、转录组学、蛋白质组学、

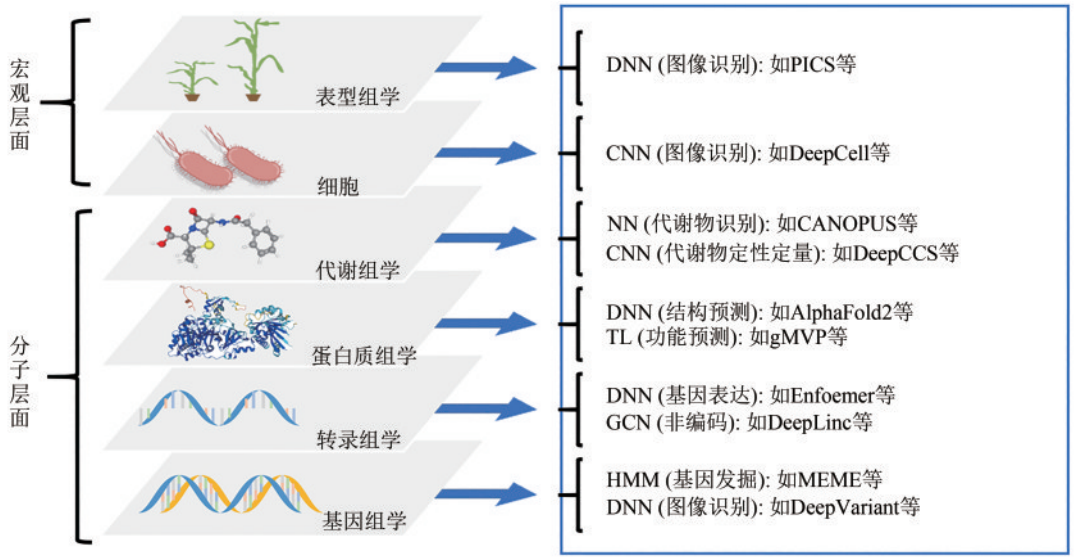


图5 序列数据的类型，以及相应的人工智能分析方法

DNN—深度神经网络；CNN—卷积神经网络；NN—神经网络；TL—迁移学习；GCN—图卷及网络；HMM—隐马尔科夫模型

Fig. 5 Types of sequence data and corresponding AI analysis methods

DNN— deep neural network; CNN— convolutional neural network; NN— neural network; TL— transfer learning; GCN— graph convolutional network; HMM— hidden markov model

代谢组学等层面，预测各种变异和调控规律；在宏观层面的细胞和表型组学层面，通过图像识别等方法进行各类样本分类[70]。随着计算机性能的不断提升，超级计算机强大的数据处理能力可以对TB级的海量基因组数据进行处理和挖掘，从而极大地缩短基因检测的时间，提高基因检测效率。将人工智能方法应用于海量的基因组数据，可以带来传统医疗向精准医疗的范式转变，人工智能方法能使医生和研究人员更准确地预测出预防与治疗方法在哪些人群中更起作用[71]。

4.2 BGC挖掘的人工智能方法：经典方法和发展趋势

伴随着生物序列人工智能分析方法能力的不断提高，BGC挖掘的方法也在不断更新换代。其中antiSMASH[22]、ClusterFinder[72]、MetaBGC[24]、DeepBGC[25]是成功应用于各领域的经典人工智能数据挖掘方法（图6）。

（1）antiSMASH工具集[22] antiSMASH在数据库基础上提供了一系列基于人工智能的计算工具，是目前寻找代谢基因簇最常用的软件之一。其主体功能主要基于的原理是：参与代谢途径中生物合成酶的基因在染色体上一般成簇排列，基于指定类型的模型，可以准确鉴定所有已知的次级代谢基因簇。在antiSMASH中，将次级代谢基因簇分为了数十类，然后通过序列比对等方法进行BGC的同源比对和发掘[73]。通过分析与目的基因相似的BGC结果，可以大致解读出目的基因的功能[74]。除此之外还提供了一些独立的工具，如由质谱引导的肽挖掘工具Pep2Path[75]、抗生素耐药性靶标搜寻器ARTS[76]和sgRNA设计工具CRISPy-web[77]等。

（2）ClusterFinder[72] ClusterFinder基于隐马尔可夫模型（hidden Markov models，HMM），它将BGC的核苷酸序列转换为一串连续的Pfam结构域，因为仅基于Pfam域频率，ClusterFinder能更精准地识别新型BGC。且有别于在此之前的算法只能识别少数BGC类别，ClusterFinder基于手动汇总的732个BGC训练集可以检测数种特征明确的基因簇类别，提供基因簇识别问题更通用的解决方案。将该算法应用到人类相关的微生物组中，鉴定出3118个小分子BGC，在临床试验中发现一类硫肽抗生素的BGC，随后通过实验确定了硫肽抗生素lactocillin的结构，并证明其对革兰氏阳性阴道病原体具有一定的抗菌活性[44]。

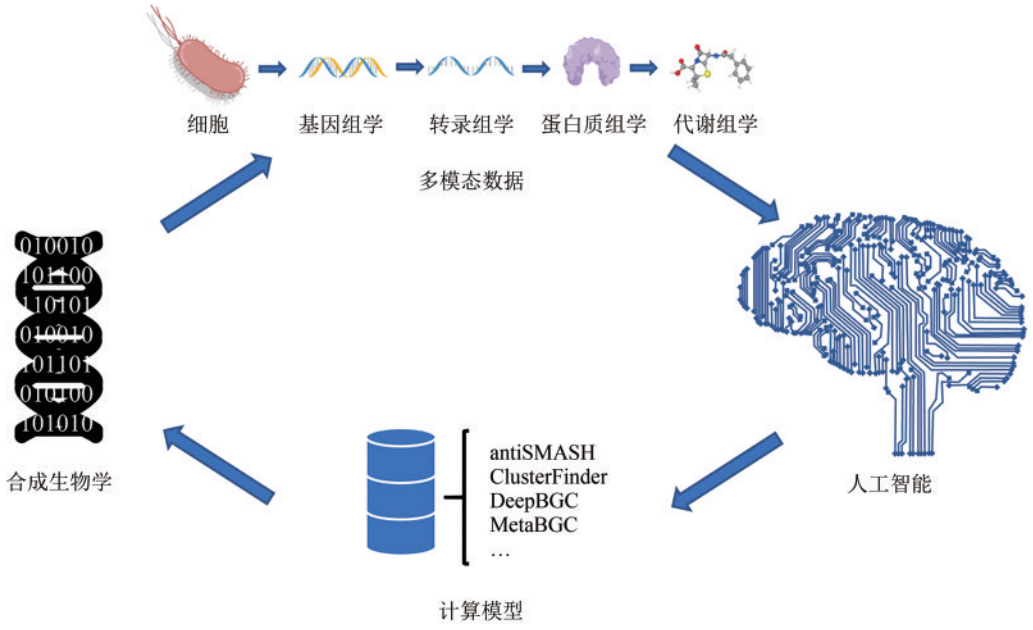


图 6　利用人工智能进行 BGC 挖掘的现状和趋势

（从数据出发，通过人工智能方法进行数据挖掘和模型构建，进而服务于合成生物学的转化研究，产生更多的多模态数据，形成良性循环）

Fig. 6　Status quo and trend of BGC mining using artificial intelligence

(Starting from the data, data mining and model construction are carried out with artificial intelligence methods, thus serving the transformation research of synthetic biology, generating more multimodal data and forming a virtuous cycle.)

（3）MetaBGC [24]　MetaBGC 方法是一种基于“读段”（reads）的算法，能够从人类微生物组中发掘之前从未被报道过的 BGC。在不需要分离培养细菌或测序的情况下，该算法允许直接在人类微生物组衍生的宏基因组测序数据中识别 BGC：通过构建基于群落画像的隐马尔可夫模型，可在单一的宏基因组读取水平上识别、定量和聚集微生物组衍生的 BGC。研究人员使用 MetaBGC 的算法在口腔、肠道和皮肤这三个部位的宏基因组样本发现了多种新型酶的 BGC，即 Ⅱ 型聚酮化合物合酶 BGC，简称为 TⅡ-PKS BGC [78-79]，并运用合成生物学策略将两种 BGC 进行异源表达，纯化与确定了产物的结构，发现其具有抗菌活性，这一结果揭示了人类微生物组产生先导化合物的能力。

（4）DeepBGC [25]　DeepBGC 使用深度学习来检测细菌和真菌基因组中的 BGC。DeepBGC 使用了双向长期短期记忆递归神经网络 [80] 和类似 word2vec [81] 的 Pfam 蛋白域嵌入，并使用随机森林分类器 [82] 预测产品类别和检测到的 BGC 的活性。将 DeepBGC 应用到实际的细菌基因组中，能预测出具有编码抗生素活性分子的全新 BGC 候选物。

发掘全新的 BGC 个例和 BGC 类型是微生物组研究中比较重要的数据挖掘目标 [83]，然而现有的数据挖掘方法难以发掘新型 BGC [84]。基于更大的 BGC 数据集构建更加智能的挖掘模型，有可能发掘新型 BGC [53]。在 BGC 数据集方面，BiG-SLiCE 方法 [53] 能将 BGC 投射到欧几里得空间，以便使用时间复杂度为近线性的分区聚类算法，有助于大型 BGC 数据集的分析。此外，Medema 等 [85] 提出的基于网络的计算框架（biosynthetic gene similarity clustering and prospecting engine， BiG-SCAPE）可用于 BGC 的聚类，以便更好地分析大数据集上微生物群落的生物合成潜力。在 BGC 挖掘模型方面，基于自然语言处理（natural language processing，NLP）技术的深度学习方法 Genomic-NLP 已经被成功地用于解码未知微生物基因的功能 [86]。在未来的研究中，开发基于 NLP 技术的人工智能模型有可能发掘出与现有数据库中已知的 BGC 不存在任何同源性，然而在代谢产物方面又有一定关系的新型 BGC。

5 新型BGC的挖掘与功能验证案例

新型BGC的功能验证，通常是通过培养实验来完成的[84]。人工智能数据挖掘（artificial intelligence data mining）和培养组学（culturomics）各自都有明显的优缺点，并且它们之间具有极强的互补性[87]（图7）。高通量测序方法能短时间内产生大量数据，再由人工智能方法迅速挖掘出有用信息；而来自于测序的数据挖掘方法，也需要由培养组学来补充未知细菌的生长条件等信息[88]。

新型BGC转化的应用范围很广，在临床、环境和生物制造方面均有非常迫切的需求[43]。目前有害生物对抗生素、癌症化疗药物和杀虫剂的耐药性上升，这一现象是现代医学与农业的主要威胁，而微生物次级代谢产物是解决这一问题的主要有效方法之一[89]，即通过发掘新型BGC合成新型次级代谢产物，从而开发出新型产品消除或减缓有害生物对人类及农作物的危害。

5.1 肠道微生物BGC的挖掘和分析研究

2019年，一项人类肠道微生物宏基因组挖掘工作揭示了未培养的细菌基因组编码数百种新的生物合成基因簇，并具有独特的功能[90]。课题组通过从11 850个人类肠道微生物群中重建92 143个宏基因组组装基因组（metagenome assembled genome，MAG），鉴定了1952个未培养的候选细菌物种。这些未经培养的细菌物种及其基因组大大扩展了人类肠道微生物群的已知物种库，将目前的系统发育多样性增加了281%。这些候选物种编码数百个新的生物合成基因簇，并在铁-硫和离子结合等代谢方面具有独特的功能，揭示了未培养肠道细菌的多样性，为肠道微生物群的分类和功能特征提供了前所未有的解决方案[91]。

5.2 土壤环境微生物BGC的挖掘和分析研究

2018年，研究人员基于草原土壤的宏基因组数据，重建了上千个基因组，其中几百个近乎完整（near-complete），并鉴定出先前未被研究过的微生物（一类酸杆菌），这些微生物能编码多种聚酮化合物和非核糖体肽合成的基因组簇[92]。具体而言，研究者鉴定出了两个来自不同谱系类群的酸杆菌（Acidobacteria）基因组，每个基因组都拥有一个异常庞大的生物合成基因库，并且含有多达15个大型聚酮化合物和非核糖体肽生物合成基因位点。为了追踪土壤中聚酮化合物合成酶和非核糖体肽合成酶基因的表达，研究者设计了一个微观操作实验（microcosm manipulation experiment），采集了120个时间点的样品，使用转录组学的手段，发现基因簇对不同环境扰动的响应情况并不相同。通过对微生物的转录共表达网络分析，发现生物

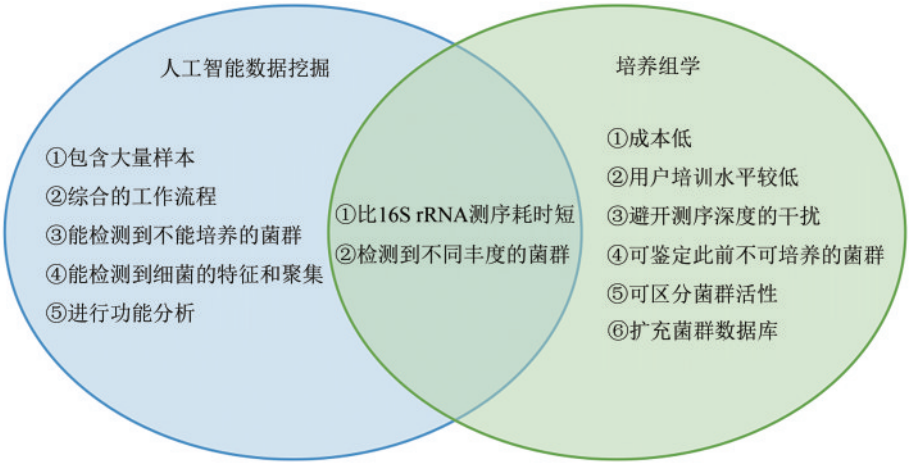


图7 人工智能数据挖掘和培养组学的各自优缺点和互补性

（相关方法优缺点的罗列，是基于互相比较和与传统分子生物学方法比较的结果）

Fig. 7 Advantages, disadvantages and complementarities of artificial intelligence data mining and culturomics

(The list of advantages and disadvantages of the relevant methods is based on the results of comparison with each other and with traditional molecular biological methods as well.)

合成基因的表达与双组分系统、转录激活、假定抗微生物剂抗性和铁调节模块的基因相关，这一结果将代谢物生物合成与环境感知和生态竞争过程联系起来。作者因此判断，土壤微生物的生物合成潜力以前被大大低估了，而这些微生物代表了一种天然产物来源，能够进行转化研究以满足人们对新型抗生素和其他先导化合物的需求。据文献报道，上述聚酮化合物和非核糖体肽生物合成基因簇来自于Acidobacteria、Verrucomicobia和Gemmatimonadetes以及候选门Rokubacteria的微生物。这些微生物类群在土壤中非常丰富，但过往研究并没有把次生代谢产物与基因组信息联系起来[93-95]。

5.3 水体环境微生物BGC的挖掘和实验验证

2022年，瑞士苏黎世联邦理工学院的研究团队借助基因组学技术和大数据挖掘方法，发现了多种海洋细菌生物合成基因簇，相关成果在*Nature*发表[32]。研究团队首先获取了全球215个采样点不同深度层共1038个海水样本的基因测序数据，构建了26 293种海洋微生物基因组，其中2790种来自新发现的细菌。结合已公布的基因组数据，研究人员创建了海洋微生物组学数据库（ocean microbiomics database， OMD），发现了39 055个生物合成基因簇，参与约6873种化合物的生物合成过程。进一步实验验证两类与任何已知BGC不相似的RiPP生物合成簇能产生新的代谢物，表明了部分基因簇在亚磷酸盐等化合物的生物合成中起着关键作用。该研究通过基因组学方法发现了新型海洋细菌和生物合成基因簇，并对部分BGC进行了实验验证，其研究成果对海洋生态、生物进化和天然产物等领域的研究具有重要意义[96]。

5.4 重要天然产物的发掘和再利用

硒（Se）是一种天然的非金属元素，主要存在于硒蛋白和硒酸生物聚合物中。由于硒具有营养学和毒理学的双重作用，因此在医学和生物学领域广受关注[97]。2022年，发表在*Nature*上的一项新研究确定了第一条将硒引入微生物的小分子生物合成途径[98]。首先，由于*SelD*基因编码了细胞内所有已知硒代谢过程的第一步，因此科研人员利用无假设的方式从美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）数据库中搜索了*SelD*的遗传背景，具体而言即通过量化了一个或多个碱基对与*selD*开放阅读框重叠的基因丰度，识别与其共定位的基因。结果表明，前5个*selD*的重叠基因包括*SelA*、*SelU*、*yedF*和*duf3343*，后两个基因被认为在硒的还原和/或转运中发挥尚未确定的作用。其次，对*SelD*–*tigr04348*遗传背景的深入研究揭示了第三种常见的共定位基因*egtB*的同源物，其编码麦角硫因生物合成中C—S键形成酶[99]。之后，科研人员通过代谢组学和生化方法表征上述生物合成途径，发现含有*SelD-egtB-tigr04348*基因簇的放线菌*Amycolatopsis palatopharyngis*和争论贪噬菌*Variovorax paradoxus*可以产生麦角硫因及其硒酮类似物。进一步分析揭示硒酮实际上是新基因簇的产物。该团队将其命名为“Sen”，*SenA*、*SenB*和*SenC*分别编码*egtB*同源物、一个假定的糖基转移酶和一个*SelD*同源物。这些发现证明SenB是一类新的硒糖合酶，继SelA和SelU之后成为迄今为止第三种C—Se键形成酶。这标志着硒元素首次在天然产物中被发现，并为硒生物学研究开辟了更广阔的前景。

5.5 天然药物资源的发掘和再利用

2022年，*Nature Catalysis*发表了丝状真菌来源萜类生物合成基因簇的高效挖掘研究工作[30]。该研究基于antiSMASH开发了“基因簇功能元件理性可控重组”策略，实现了萜类沉默基因簇的批量挖掘及高效合成，有效解决了困扰该研究领域的“三低”（研究通量低、产物集中度低、产量低）研究瓶颈[100]，显著提高了活性新产物的合成效率。该研究借助自动化平台实现了丝状真菌来源萜类基因簇及其产物的高通量挖掘，并开发了真菌高效萜类前体供给底盘，实现了产物的高效合成。在丝状真菌米曲霉（*Aspergillus oryzae*）底盘中，通过模块化组合重构了5种真菌来源的39个Ⅰ型萜类生物合成基因簇，随后借助抗炎活性高通量筛选模型快速锁定高活性产物及其对应

的突变株，紧接着回溯突变株对应的基因簇，解析了具有显著抗炎活性的二倍半萜化合物mangicol类（酯萜多元醇）[101]的生物合成机理。

以上应用案例表明：针对微生物组的BGC挖掘和解读，能够极大地提高天然产物的发掘效率，促进生物工程与合成生物学的发展，并在多领域取得明显的成效。

6 结论和展望

本文通过对BGC相关微生物组大数据以及相关数据挖掘方法的介绍，配合详实的案例，描绘了BGC挖掘和转化研究方面的全景图。首先，较全面地回顾了BGC挖掘的意义和瓶颈问题，指出当前实验验证的BGC数据不够全面，而基于序列比对的BGC挖掘难以发现新类型的BGC资源。其次，系统性地总结了当前BGC发掘中的数据资源和挖掘方法，尤其是人工智能方法，指出了其巨大的潜力。同时，通过回顾当前培养组学和合成生物学方面的技术进展，指出了干湿结合方法对于验证新发掘的BGC的重要价值。最后，通过展示到表型的新发掘BGC案例，指出BGC挖掘被应用于不同的研究领域，且取得了较好的研究成果。

6.1 BGC挖掘的研究是合成生物学与人工智能交叉研究方向上非常重要的一环

BGC挖掘的研究，是合成生物学与人工智能交叉研究方向上非常重要的一个环节，其重要性体现在方法上代表着人工智能生物数据挖掘的趋势，在转化应用上也具有非常高的价值。

首先，BGC挖掘的研究，是合成生物学中重要的一个部分。合成生物学（synthetic biology）是一门汇集生物学、基因组学、工程学和信息学等多种学科的交叉学科，其实现的技术路径是运用系统生物学和工程学原理，以基因组和生化分子合成为基础，综合生物化学、生物物理和生物信息等技术，旨在设计、改造、重建生物分子、生物元件和生物分化过程，以构建具有生命活性的生物系统[102]。将新型BGC作为原件，对已有的底盘生物进行理性设计，是合成生物学的典型应用场景[103]，而利用生物信息学分析和计算工具，能挖掘大量未知的BGC，再将这些BGC通过上述合成生物学手段进行验证，即完成BGC的挖掘研究，这一流程将极大地加快天然产物的开发与利用。

其次，BGC挖掘的研究，在方法上代表着较为高级的生物大数据挖掘趋势[84]：BGC在序列和功能上的特征决定了针对其挖掘的人工智能手段必须比传统的单个基因挖掘方法要复杂，这种需要上下文感知的人工智能挖掘手段，是生物大数据挖掘趋势[86]。人工智能与合成生物学的结合，可以实现更为智能化、数字化、工程化的合理设计和优化，这也是BGC挖掘研究的重点和难点[84]。另外需要指出的是：人工智能与合成生物学的结合，干湿实验的结合，都指向更为高通量的“发掘-验证”流程，而高通量的“发掘-验证”流程，能够更为快速地发掘潜在新类型的BGC并加以验证，具有明显的工程属性，同时也能够极大地提高BGC发掘、验证和转化的效率。

最后，BGC挖掘的研究，在转化应用上具有非常高的价值[9]：通过人工智能挖掘的元件和模块，可以直接结合合成生物学的研究进行验证[104]，并快速进行转化应用，尤其是在精准医学等转化领域日益精进的今天，通过有效开展BGC及其相关化合物的转化研究，快速有效地实现从实验室到临床（from bench to bedside）的转化，具有非常高的经济价值和社会价值[105]。

6.2 BGC挖掘的研究需要重视干湿实验等方面全方位结合

此外，BGC挖掘研究的成功，十分依赖于BGC数据库和相关基因实体库相结合，依赖于人工智能挖掘和培养实验验证相结合。只有在干湿实验等方面全方位结合，才能更有效地实现BGC挖掘、验证以及转化等方面的研究。

BGC数据库和相关基因实体库相结合，能够更好地推动BGC挖掘研究和转化应用的开展[6]，是保证BGC挖掘研究和转化应用顺利开展的基本材料和数据条件[106]。基于数据库的不断更新，配

合相关序列和结构等规律的发掘，为更全面的BGC发掘打下了数据基础。同时实体库能较为便捷地进行BGC验证实验，也为发掘新型BGC提供了保障。因此，作为数据基础的数据库和实体库相结合，能够更好地推动BGC挖掘研究和转化应用的开展。

人工智能挖掘和培养实验验证相结合，是保证BGC挖掘研究和转化应用顺利开展的基本技术条件[25]。传统的基于序列比对的BGC挖掘难以发现新类型的BGC资源，而利用人工智能技术，基于已有的BGC及其同源序列集合进行大数据建模，将有望批量发掘新型BGC。另一方面，培养组学等实验技术，将能够快速有效地验证新发掘BGC的有效性。因此，人工智能挖掘和培养实验验证技术作为关键引擎，是保证BGC挖掘研究和转化应用顺利开展的基本技术条件。

由上述讨论可知，BGC在系统生物学与合成生物学中具有核心地位（图8）：不但在数据上打通了数据库和实体库，而且在技术上打通了人工智能挖掘和培养实验验证。因此BGC的研究能够紧密连接系统生物学与合成生物学，实现从数据到模型，从验证到应用的无缝转化。

系统生物学和合成生物学协同发展的趋势，尤其是作为在系统生物学与合成生物学中具有核心地位之一的BGC挖掘与转化研究快速发展的趋势，会更为突出地显示出来。而多组学技术和人工智能分析方法，将会极大地助力这一方向的快速进步。我们乐观地展望，在BGC被充分挖掘和认识之后，系统生物学与合成生物学的结合将会深刻地改变世界：从科学探索方面来说，新发掘的BGC能够快速地被研究并转化于实际应用，高效实现各类小分子化合物从“实验到临床”（from bench to bedside）；从健康和环境领域等方面来说，从需求端倒推BGC资源的特征，能够快速地实现转化研究领域中需要的特定功能小分子合成系统的“即插即用”（plug-and-play）。从而在技术上较为高效、准确、完整、安全地实现针对BGC合成生物系统从理解到创造（from understanding to creation）的过程。

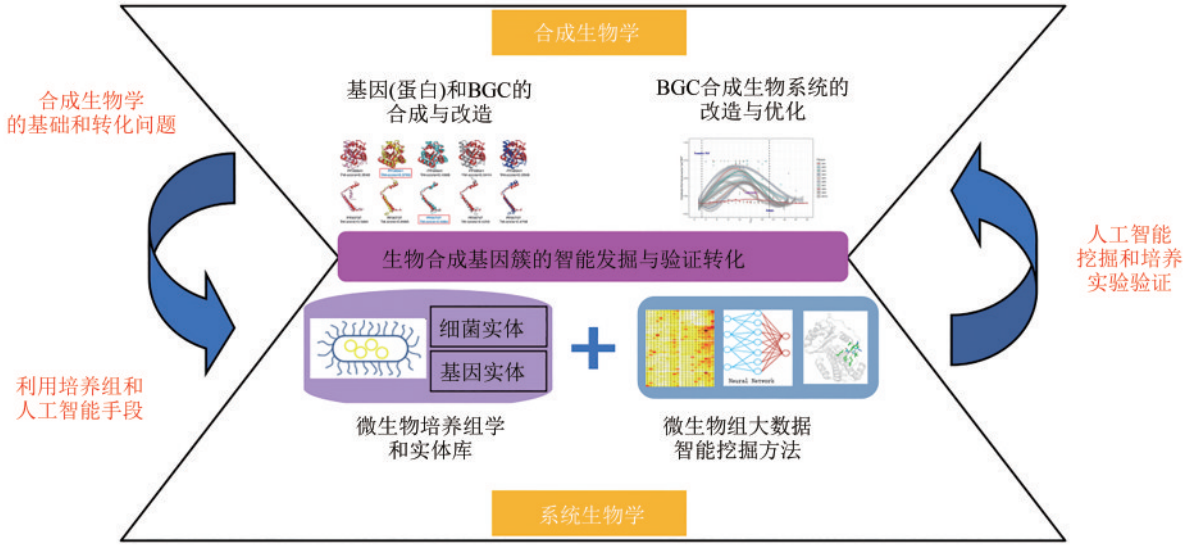


图8　BGC在系统生物学与合成生物学中的核心地位

（生物合成基因簇的智能发掘与验证转化的研究，不但在数据上打通了数据库和实体库，而且在技术上打通了人工智能挖掘和培养实验验证。生物合成基因簇的智能发掘与验证转化的研究，能够紧密连接系统生物学与合成生物学，实现从数据到模型、从验证到应用的无缝转化）

Fig. 8　BGC's central role in systems biology and synthetic biology

(Research on intelligent mining and verification transformation of biosynthetic gene clusters not only connects BGC database with entity database, but also connects artificial intelligence mining and culture experiment verification. Research on intelligent discovery and transformation verification for biosynthetic gene clusters can closely link systems biology and synthetic biology, and realize seamless transformation from data to model and from verification to application.)

参 考 文 献

[1] ZHANG L X, DEMAIN A L. Natural Products: Drug Discovery, and Therapeutic Medicines[M]. Clifton, New Jersey: Humana Totowa Press, 2005: 382.

[2] SANCHEZ S, GUZMÁN-TRAMPE S, ÁVALOS M, et al. Microbial natural products[M/OL]//Natural Products in Chemical Biology. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2012: 65-108[2022-12-01]. https://onlinelibrary. wiley. com/doi/10.1002/9781118391815.ch3.

[3] LLOYD K G, STEEN A D, LADAU J, et al. Phylogenetically novel uncultured microbial cells dominate earth microbiomes[J]. mSystems, 2018, 3(5): e00055-18.

[4] WOODRUFF H B. Natural products from microorganisms[J]. Science, 1980, 208(4449): 1225-1229.

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

[6] WASCHULIN V, BORSETTO C, JAMES R, et al. Biosynthetic potential of uncultured Antarctic soil bacteria revealed through long-read metagenomic sequencing[J]. The ISME Journal, 2022, 16(1): 101-111.

[7] MARTINET L, NAÔMÉ A, DEFLANDRE B, et al. A single biosynthetic gene cluster is responsible for the production of bagremycin antibiotics and ferroverdin iron chelators[J]. mBio, 2019, 10(4): e01230-19.

[8] SAWANT A M, VAMKUDOTH K R. Biosynthetic process and strain improvement approaches for industrial penicillin production[J]. Biotechnology Letters, 2022, 44(2): 179-192.

[9] KWON M J, STEINIGER C, CAIRNS T C, et al. Beyond the biosynthetic gene cluster paradigm: genome-wide coexpression networks connect clustered and unclustered transcription factors to secondary metabolic pathways[J]. Microbiology Spectrum, 2021, 9(2): e00898-21.

[10] MARTÍN J F. Molecular control of expression of penicillin biosynthesis genes in fungi: regulatory proteins interact with a bidirectional promoter region[J]. Journal of Bacteriology, 2000, 182(9): 2355-2362.

[11] MILLER B L, MILLER K Y, ROBERTI K A, et al. Position-dependent and-independent mechanisms regulate cell-specific expression of the *SpoC1* gene cluster of *Aspergillus nidulans*[J]. Molecular and Cellular Biology, 1987, 7(1): 427-434.

[12] DEMAIN A L, FANG A. The natural functions of secondary metabolites[M]//Advances in Biochemical Engineering/Biotechnology: History of modern biotechnology Ⅰ, Berlin: Springer, 2000, 69: 1-39.

[13] 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.

[14] ARNISON P G, BIBB M J, BIERBAUM G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature[J]. Natural Product Reports, 2013, 30(1): 108-160.

[15] ZHONG Z, HE B B, LI J, et al. Challenges and advances in genome mining of ribosomally synthesized and post-translationally modified peptides (RiPPs) [J]. Synthetic and Systems Biotechnology, 2020, 5(3): 155-172.

[16] 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.

[17] EPSTEIN S C, CHARKOUDIAN L K, MEDEMA M H. A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences[J]. Standards in Genomic Sciences, 2018, 13: 16.

[18] 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.

[19] TERLOUW B R, BLIN K, NAVARRO-MUÑOZ J C, et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters[J]. Nucleic Acids Research, 2023, 51(D1): D603-D610.

[20] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215 (3): 403-410.

[21] RABINER L, JUANG B. An introduction to hidden Markov models[J]. IEEE ASSP Magazine, 1986, 3(1): 4-16.

[22] BLIN K, SHAW S, KLOOSTERMAN A M, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities[J]. Nucleic Acids Research, 2021, 49(W1): W29-W35.

[23] MOHIMANI H, KERSTEN R D, LIU W T, et al. Automated genome mining of ribosomal peptide natural products[J]. ACS Chemical Biology, 2014, 9(7): 1545-1551.

[24] SUGIMOTO Y, CAMACHO F R, WANG S, et al. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome[J]. Science, 2019, 366(6471): eaax9176.

[25] HANNIGAN G D, PRIHODA D, PALICKA A, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction[J]. Nucleic Acids Research, 2019, 47(18): e110.

[26] RUIZ B, CHÁVEZ A, FORERO A, et al. Production of microbial secondary metabolites: regulation by the carbon source[J]. Critical Reviews in Microbiology, 2010, 36(2): 146-167.

[27] O'BRIEN J, WRIGHT G D. An ecological perspective of microbial secondary metabolism[J]. Current Opinion in Biotechnology, 2011, 22(4): 552-558.

[28] SEYEDSAYAMDOST M R. Toward a global picture of bacterial secondary metabolism[J]. Journal of Industrial Microbiology & Biotechnology, 2019, 46(3/4): 301-311.

[29] KALKREUTER E, PAN G H, CEPEDA A J, et al. Targeting bacterial genomes for natural product discovery[J]. Trends in Pharmacological Sciences, 2020, 41(1): 13-26.

[30] YUAN Y J, CHENG S, BIAN G K, et al. Efficient exploration of terpenoid biosynthetic gene clusters in filamentous fungi[J]. Nature Catalysis, 2022, 5(4): 277-287.

[31] BURIAN J, LIBIS V K, HERNANDEZ Y A, et al. High-throughput retrieval of target sequences from complex clone libraries using CRISPRi[J]. Nature Biotechnology, 2022: 1-5.

[32] PAOLI L, RUSCHEWEYH H J, FORNERIS C C, et al. Biosynthetic potential of the global ocean microbiome[J]. Nature, 2022, 607(7917): 111-118.

[33] PATEL J R, OH J S, WANG S Q, et al. Cross-kingdom expression of synthetic genetic elements promotes discovery of metabolites in the human microbiome[J]. Cell, 2022, 185(9): 1487-1505.e14.

[34] DIRENÇ MUNGAN M, BLIN K, ZIEMERT N. ARTS-DB: a database for antibiotic resistant targets[J]. Nucleic Acids Research, 2022, 50(D1): D736-D740.

[35] NAYFACH S, ROUX S, SESHADRI R, et al. A genomic catalog of Earth's microbiomes[J]. Nature Biotechnology, 2021, 39 (4): 499-509.

[36] VAN BERGEIJK D A, TERLOUW B R, MEDEMA M H, et al. Ecology and genomics of Actinobacteria: new concepts for natural product discovery[J]. Nature Reviews Microbiology, 2020, 18(10): 546-558.

[37] BARBOUR A, WESCOMBE P, SMITH L. Evolution of lantibiotic salivaricins: new weapons to fight infectious diseases[J]. Trends in Microbiology, 2020, 28(7): 578-593.

[38] CARRIÓN V J, PEREZ-JARAMILLO J, CORDOVEZ V, et al. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome[J]. Science, 2019, 366 (6465): 606-612.

[39] ZHAO H, FU S L, YU Y F, et al. *MetaMed*: Linking microbiota functions with medicine therapeutics[J]. mSystems, 2019, 4 (5): e00413-19.

[40] CHU J, VILA-FARRES X, BRADY S F. Bioactive synthetic-bioinformatic natural product cyclic peptides inspired by nonribosomal peptide synthetase gene clusters from the human microbiome[J]. Journal of the American Chemical Society, 2019, 141(40): 15737-15741.

[41] WANG L L, RAVICHANDRAN V, YIN Y L, et al. Natural products from mammalian gut microbiota[J]. Trends in Biotechnology, 2019, 37(5): 492-504.

[42] SKELLAM E. Strategies for engineering natural product biosynthesis in fungi[J]. Trends in Biotechnology, 2019, 37(4): 416-427.

[43] RUTLEDGE P J, CHALLIS G L. Discovery of microbial natural products by activation of silent biosynthetic gene clusters[J]. Nature Reviews Microbiology, 2015, 13(8): 509-523.

[44] DONIA M S, CIMERMANCIC P, SCHULZE C J, et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics[J]. Cell, 2014, 158(6): 1402-1414.

[45] GOLAKOTI T, YOSHIDA W Y, CHAGANTY S, et al. Isolation and structure determination of nostocyclopeptides A1 and A2 from the terrestrial Cyanobacterium *Nostoc* sp. ATCC53789[J]. Journal of Natural Products, 2001, 64(1): 54-59.

[46] HAMMAMI R, ZOUHIR A, LE LAY C, et al. BACTIBASE second release: a database and tool platform for bacteriocin characterization[J]. BMC Microbiology, 2010, 10: 22.

[47] KAUTSAR S A, BLIN K, SHAW S, et al. BiG-FAM: the biosynthetic gene cluster families database[J]. Nucleic Acids Research, 2021, 49(D1): D490-D497.

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

[49] 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.

[50] 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.

[51] PALANIAPPAN K, CHEN I M A, CHU K, et al. IMG-ABC v.5.0: an update to the IMG/Atlas of biosynthetic gene clusters knowledgebase[J]. Nucleic Acids Research, 2020, 48(D1): D422-D430.

[52] O'BRIEN R V, DAVIS R W, KHOSLA C, et al. Computational identification and analysis of orphan assembly-line polyketide synthases[J]. The Journal of Antibiotics, 2014, 67(1): 89-97.

[53] KAUTSAR S A, VAN DER HOOFT J J J, DE RIDDER D, et al. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters[J]. GigaScience, 2021, 10 (1): giaa154.

[54] TRAN P N, YEN M R, CHIANG C Y, et al. Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi[J]. Applied Microbiology and Biotechnology, 2019, 103(8): 3277-3287.

[55] 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 (suppl_2): W339-W346.

[56] MISTRY J, FINN R D, EDDY S R, et al. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions[J]. Nucleic Acids Research, 2013, 41(12): e121.

[57] EDDY S R. Profile hidden Markov models[J]. Bioinformatics, 1998, 14(9): 755-763.

[58] KJÆRBØLLING I, MORTENSEN U H, VESTH T, et al. Strategies to establish the link between biosynthetic gene clusters and secondary metabolites[J]. Fungal Genetics and Biology, 2019, 130: 107-121.

[59] RABE P, KAMPS J J A G, SUTHERLIN K D, et al. X-ray free-electron laser studies reveal correlated motion during isopenicillin *N* synthase catalysis[J]. Science Advances, 2021, 7(34): eabh0250.

[60] CHING T, HIMMELSTEIN D S, BEAULIEU-JONES B K, et al. Opportunities and obstacles for deep learning in biology and medicine[J]. Journal of the Royal Society, Interface, 2018, 15(141): 20170387.

[61] JING Y K, BIAN Y M, HU Z H, et al. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era[J]. The AAPS Journal, 2018, 20(3): 58.

[62] SHEN D G, WU G R, SUK H I. Deep learning in medical image analysis[J]. Annual Review of Biomedical Engineering, 2017, 19: 221-248.

[63] HAMET P, TREMBLAY J. Artificial intelligence in medicine[J]. Metabolism 2017, 69S: S36-S40.

[64] JURTZ V I, JOHANSEN A R, NIELSEN M, et al. An introduction to deep learning on biological sequence data: examples and solutions[J]. Bioinformatics, 2017, 33(22): 3685-3690.

[65] KANDEL M E, HE Y R, LEE Y J, et al. Phase imaging with computational specificity (PICS) for measuring dry mass changes in sub-cellular compartments[J]. Nature Communications, 2020, 11: 6256.

[66] BANNON D, MOEN E, SCHWARTZ M, et al. DeepCell Kiosk: scaling deep learning—enabled cellular image analysis with Kubernetes[J]. Nature Methods, 2021, 18(1): 43-45.

[67] AVSEC Ž, AGARWAL V, VISENTIN D, et al. Effective gene expression prediction from sequence by integrating long-range interactions[J]. Nature Methods, 2021, 18(10): 1196-1203.

[68] LI R Z, YANG X R. *De novo* reconstruction of cell interaction landscapes from single-cell spatial transcriptome data with DeepLinc[J]. Genome Biology, 2022, 23(1): 124.

[69] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596 (7873): 583-589.

[70] CUI M, ZHANG D Y. Artificial intelligence and computational pathology[J]. Laboratory Investigation, 2021, 101(4): 412-422.

[71] MESKO B. The role of artificial intelligence in precision medicine[J]. Expert Review of Precision Medicine and Drug Development, 2017, 2(5): 239-241.

[72] 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.

[73] MINOWA Y, ARAKI M, KANEHISA M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes[J]. Journal of Molecular Biology, 2007, 368(5): 1500-1517.

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

[75] MEDEMA M H, PAALVAST Y, NGUYEN D D, et al. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products[J]. PLoS Computational Biology, 2014, 10(9): e1003822.

[76] ALANJARY M, KRONMILLER B, ADAMEK M, et al. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery [J]. Nucleic Acids Research, 2017, 45(W1): W42-W48.

[77] BLIN K, PEDERSEN L E, WEBER T, et al. CRISPy-web: an online resource to design sgRNAs for CRISPR applications[J]. Synthetic and Systems Biotechnology, 2016, 1(2): 118-121.

[78] HERTWECK C, LUZHETSKYY A, REBETS Y, et al. Type Ⅱ polyketide synthases: gaining a deeper insight into enzymatic teamwork [J]. Natural product reports, 2007, 24(1): 162-190.

[79] FENG Z Y, 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.

[80] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5/6): 602-610.

[81] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. arXiv, 2013: 1301.3781[2022-12-01]. https://arxiv.org/abs/1301.3781.

[82] BREIMAN L. Random forests[J].Machine Learning, 2001, 45 (1): 5-32.

[83] SCHERLACH K, HERTWECK C. Mining and unearthing hidden biosynthetic potential[J]. Nature Communications, 2021, 12(1): 3864.

[84] ALBARANO L, ESPOSITO R, RUOCCO N, et al. Genome mining as new challenge in natural products discovery[J]. Marine Drugs, 2020, 18(4): 199.

[85] NAVARRO-MUÑOZ J C, SELEM-MOJICA N, MULLOWNEY M W, et al. A computational framework to explore large-scale biosynthetic diversity[J]. Nature Chemical Biology, 2020, 16 (1): 60-68.

[86] MILLER D, STERN A, BURSTEIN D. Deciphering microbial gene function using natural language processing[J]. Nature Communications, 2022, 13: 5731.

[87] HA C W Y, DEVKOTA S. The new microbiology: cultivating

the future of microbiome-directed medicine[J]. American Journal of Physiology Gastrointestinal and Liver Physiology, 2020, 319(6): G639-G645.

[88] LAGIER J C, DUBOURG G, MILLION M, et al. Culturing the human microbiota and culturomics[J]. Nature Reviews Microbiology, 2018, 16(9): 540-550.

[89] DEMAIN A L, SANCHEZ S. Microbial drug discovery: 80 years of progress[J]. The Journal of Antibiotics, 2009, 62(1): 5-16.

[90] ALMEIDA A, MITCHELL A L, BOLAND M, et al. A new genomic blueprint of the human gut microbiota[J]. Nature, 2019, 568(7753): 499-504.

[91] PASOLLI E, ASNICAR F, MANARA S, et al. Extensive unexplored human microbiome diversity revealed by over 150, 000 genomes from metagenomes spanning age, geography, and lifestyle[J]. Cell, 2019, 176(3): 649-662.e20.

[92] CRITS-CHRISTOPH A, DIAMOND S, BUTTERFIELD C N, et al. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis[J]. Nature, 2018, 558(7710): 440-444.

[93] FIERER N. Embracing the unknown: disentangling the complexities of the soil microbiome[J]. Nature Reviews Microbiology, 2017, 15(10): 579-590.

[94] BERGMANN G T, BATES S T, EILERS K G, et al. The under-recognized dominance of *Verrucomicrobia* in soil bacterial communities[J]. Soil Biology and Biochemistry, 2011, 43(7): 1450-1455.

[95] KIELAK A M, BARRETO C C, KOWALCHUK G A, et al. The ecology of acidobacteria: moving beyond genes and genomes[J]. Frontiers in Microbiology, 2016, 7: 744.

[96] MORAN M A, KUJAWINSKI E B, SCHROER W F, et al. Microbial metabolites in the marine carbon cycle[J]. Nature Microbiology, 2022, 7(4): 508-523.

[97] REICH H J, HONDAL R J. Why nature chose selenium[J]. ACS Chemical Biology, 2016, 11(4): 821-841.

[98] KAYROUZ C M, HUANG J, HAUSER N, et al. Biosynthesis of selenium-containing small molecules in diverse microorganisms[J]. Nature, 2022, 610(7930): 199-204.

[99] GONCHARENKO K V, VIT A, BLANKENFELDT W, et al. Structure of the sulfoxide synthase EgtB from the ergothioneine biosynthetic pathway[J]. Angewandte Chemie International Edition, 2015, 54(9): 2821-2824.

[100] BIAN G K, DENG Z X, LIU T G. Strategies for terpenoid overproduction and new terpenoid discovery[J]. Current Opinion in Biotechnology, 2017, 48: 234-241.

[101] RENNER M K, JENSEN P R, FENICAL W. Mangicols: structures and biosynthesis of a new class of sesterterpene polyols from a marine fungus of the genus Fusarium[J]. The Journal of Organic Chemistry, 2000, 65(16): 4843-4852.

[102] HEINEMANN M, PANKE S. Synthetic biology—putting engineering into biology[J]. Bioinformatics, 2006, 22(22): 2790-2799.

[103] LI L, LIU X C, JIANG W H, et al. Recent advances in synthetic biology approaches to optimize production of bioactive natural products in Actinobacteria[J]. Frontiers in Microbiology, 2019, 10: 2467.

[104] MALICO A A, NICHOLS L, WILLIAMS G J. Synthetic biology enabling access to designer polyketides[J]. Current Opinion in Chemical Biology, 2020, 58: 45-53.

[105] PYE C R, BERTIN M J, LOKEY R S, et al. Retrospective analysis of natural products provides insights for future discovery trends[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(22): 5601-5606.

[106] LEE N M, HWANG S K, KIM J H, et al. Mini review: genome mining approaches for the identification of secondary metabolite biosynthetic gene clusters in *Streptomyces*[J]. Computational and Structural Biotechnology Journal, 2020, 18: 1548-1556.

通讯作者： 宁康（1979—），男，教授，博士生导师。研究方向为生物信息学，微生物组学，人工智能生物学。

E-mail：ningkang@hust.edu.cn

通讯作者： 白虹（1978—），女，正高级工程师。研究方向为天然产物化学，微生物学。

E-mail：baihong@hust.edu.cn

第一作者： 赖奇龙（2000—），男，学士。研究方向为生物信息学，人工智能生物学。

E-mail：laiqilong@hust.edu.cn