微生物学通报
Microbiology China
tongbao@im.ac.cn

Oct. 20, 2013, 40(10): 1909–1919

专论与综述

对链霉菌组学数据进行深度挖掘时, 生物信息学工具和二级数据库在开展菌株特有的基因组岛和次生代谢物生物合成基因簇的识别及功能解析中有重要作用。

欧竑宇

链霉菌基因组岛和次生代谢物合成相关的生物信息学工具及数据库

欧竑宇

(上海交通大学 微生物代谢国家重点实验室 上海 200030)

摘 要: 随着 DNA 测序技术的进步, 迄今为止已有 12 个链霉菌基因组被测序。面对海量组学的数据, 急需采用生物信息学方法来大规模深度挖掘这些重要微生物资源, 进而实现链霉菌资源挖掘和代谢潜力释放的深度互动。围绕链霉菌基因组比较分析中菌株特有的基因组岛和次生代谢物生物合成基因簇的识别及功能解析等两个常见问题, 本文收集了近期开发的一些常用生物信息学工具和二级数据库。以链霉菌染色体核心区和两臂的划分、天蓝色链霉菌和变铅青链霉菌基因组岛的识别、卡特利链霉菌巨型质粒的鉴别为例, 简介了这些生物信息学资源的使用方法。此外, 还简述了我们课题组进行放线菌型整合性接合元件识别和开发硫肽生物合成基因簇预测新工具的一些尝试。生物信息学工具和二级数据库在链霉菌基因组比较分析中有重要作用, 可将研究重点迅速地聚焦在某株菌的可移动遗传元件和次生代谢物生成基因簇上, 确定其对应的菌株特有表型, 及解析新型化合物生物合成和调控机理。

关键词: 生物信息学, 链霉菌基因组岛, 整合性接合元件, 次生代谢物生物合成, 硫肽基因簇

基金项目: 国家 973 计划项目(No. 2012CB721002); 国家自然科学基金项目(No. 31371261)
***通讯作者:** Tel: 86-21-62932943; ✉: hyou@sjtu.edu.cn
收稿日期: 2013-02-15; **接受日期:** 2013-04-15

Bioinformatics tools and databases focused on genomic islands and secondary metabolite biosynthesis of *Streptomyces*

OU Hong-Yu

(*State Key Laboratory of Microbial Metabolism*, *Shanghai Jiao Tong University*, *Shanghai* 200030, *China*)

Abstract: With many *Streptomyces* genomic sequences made available by next generation sequencing (NGS), the bioinformatics resource comes as the answer to an acute demand of the big data mining. In this study, I collected the frequently used tools and open-access databases for genome alignment and island detection, suitable for the large *Streptomyces* genomes with high G+C content. To illustrate these user-friendly tools, I identified the Actinobacteria integrative and conjugative elements in 12 completely sequenced *Streptomyces* genomes, which were found to be inserted into the core regions of the *Streptomyces* chromosomes. The bioinformatics resources currently available for secondary metabolites isolated from *Streptomyces* species are also described briefly, such as the web-based tool ThioFinder to rapidly identify thiopeptide biosynthetic gene cluster from DNA sequence using a profile Hidden Markov Model approach.

Keywords: Bioinformatics, *Streptomyces* genomic islands, Actinobacteria integrative and conjugative elements, Secondary metabolite biosynthesis, Thiopeptide gene clusters

链霉菌属细菌(*Streptomyces*)是一类革兰氏阳性、多细胞的丝状土壤微生物, 是研究微生物形态分化和化学多样性的良好研究材料[1]。链霉菌具有复杂的生活周期和次生代谢途径, 产生大量具有重要价值的天然代谢物, 如大约 4 000 种以上抗生素(占自然界已知抗生素的 70%)。链霉菌形态分化、抗生素及色素合成等次生代谢过程受到其基因组遗传多样性的影响。链霉菌基因组大小约为 8−10 Mb, G+C 含量高达 70%。其染色体大部分呈线状, 以复制起点 *ori* 为中心, 可分为核心区(Core region)和两臂区(Two arm regions)。染色体核心区在链霉菌属中相对保守, 主要含有许多必需基因。而两臂及其末端不稳定, 存在大量反向重复序列, 受到基因水平转移的影响。

以 454、Solexa 和 SOLiD 为代表的第二代 DNA 测序技术的出现, 使得细菌全基因组测序工作量和花费得以锐减。迄今为止已有十多个链霉菌的遗传学蓝图通过基因组测序被揭示出来。此外, 二代测序技术的广泛应用更是推进了基于 RNA-seq 的链霉菌转录组研究。面对各种组学研究产生的大数据(Big data)剧增, 急需生物信息学工具和数据库来大规模深度挖掘这些重要微生物资源, 迅速实现菌株特有基因型(Genotype)到表型(Phenotype)的功能解析。

近年来, 生物信息学在线工具即网上服务数目已较多, 其应用也遍及链霉菌研究的各个

领域，如，能够快速预测多种次生代谢物基因簇的在线工具 antiSMASH[2]。此外，针对某一链霉菌研究课题构建的二级数据库具有专业性强和数据质量高的特点，对多种来源的不同数据进行了有效整合、人工校验和分类存储，还提供了网页浏览、关键词查询和序列比对等等面向实验人员的易用服务。如, StreptomeDB 数据库将菌株、天然化合物、生物活性和生物合成基因等信息综合在一起[3]。

本文主要围绕链霉菌基因组比较分析中常见的两个问题，菌株特有 DNA 大片段和次生代谢物生物合成基因簇的识别和功能解析，收集了一些常用生物信息学工具和开放的二级数据库。此外，以放线菌型整合性接合元件识别和硫肽生物合成基因簇预测两个问题为例，简述了我们课题组在开发生物信息学新工具和二级数据库的一些尝试。

1 链霉菌菌株特有基因组岛的识别

1.1 链霉菌基因组序列的比对

链霉菌全基因组测序相对于其他细菌难度较大，主要是因为链霉菌基因组有两个明显特点: (1) 基因组大, G+C 含量高; (2) 染色体常为线性，两臂不稳定，末端序列需要重测序。目前，已有以下12条链霉菌染色体完成全测序或拼接: *S. griseus* NBRC 13350 (NC_010572), *S. avermitilis* MA-4680 (NC_003155), *S. scabiei* 87.22 (NC_013929), *S. bingchenggensis* BCW-1 (NC_016582), *S. flavogriseus* ATCC 33331 (NC_016114), *S. venezuelae* ATCC 10712 (NC_018750), *S. violaceusniger* Tu 4113 (NC_015957), *Streptomyces* sp. SirexAA-E (NC_015953), *S. cattleya* DSM 46488 (NC_017586), *S. hygroscopicus* jinggangensis 5008 (NC_017765), *S. coelicolor* A3(2) (NC_003888), *S. lividans* TK24 (NZ_GG657756)。已测序链霉菌染色体大部分为线性结构; 只有 *S. violaceusniger* Tu 4113染色体是环形的，其复制起始位点(*ori*)位于2 kb 处。天蓝色链霉菌A3(2)染色体长度为8.7 Mb; 其中，核心区为4.9 Mb，左臂为1.5 Mb，右臂为2.3 Mb[4]。以天蓝色链霉菌为参照，我们使用 MUMmer 软件(表1)确定了其它已测序链霉菌染色体的核心区及两臂，结果如图1所示。

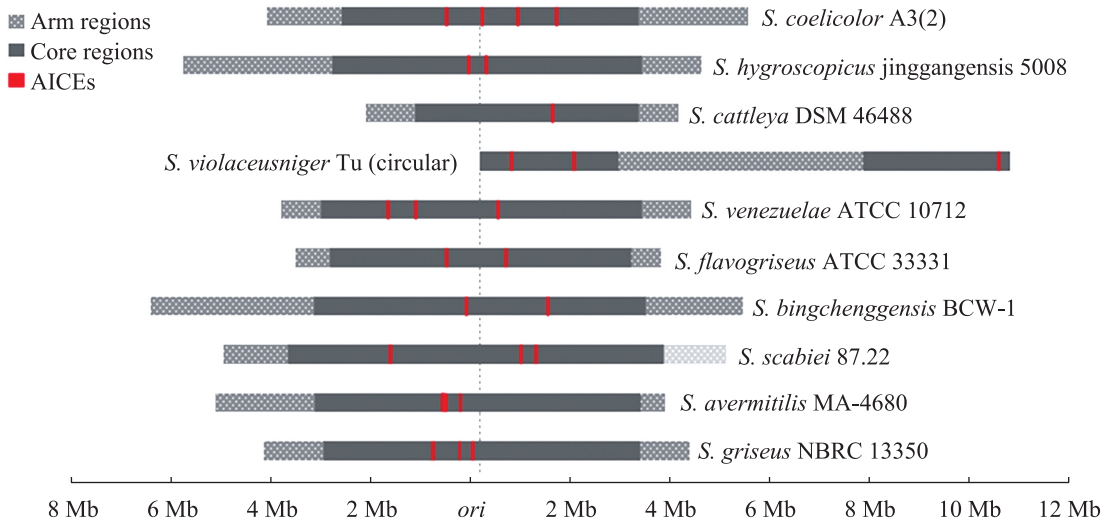

图 1 已全测序的链霉菌染色体核心区中放线菌型整合性接合元件(AICE)的分布

Fig. 1 Distribution of identified AICEs in the core regions of the completely sequenced *Streptomyces* chromosomes

我们对天然产含氟化合物的卡特利链霉菌全基因组测序后, 惊讶地发现该菌具特殊的基因组结构: 两个大复制子, 一条 6.2 Mb 线性染色体和一个 1.8 Mb 线性“巨型质粒”。脉冲场凝胶电泳实验证实了两个线性大复制子的存在。利用 mGenomeSubtractor 基因比对工具(表 1), 在巨型质粒上没有发现链霉菌染色体的保守基因。采用 MEGA 5 系统发育软件分析发现, 卡特利链霉菌质粒中与分配系统相关的 Par 蛋白(SCATT_p08020 和 SCATT_p08030)与其他链霉菌质粒 Par 蛋白发生了聚类; 而卡特利链霉菌染色体的 Par 蛋白(SCATT_09680)与其他链霉菌染色体的 Par 蛋白聚类在一起。这些分析结果支持了将 1.8 Mb 复制子称为质粒的假设。有趣的问题产生了, 巨型质粒从哪里来? 它如何与染色体互作以保持遗传稳定? 此外, 通过基因敲除实验, 我们发现了卡特利链霉菌中参与含氟化合物合成的关键基因散落于染色体和巨型质粒上[5]; 这一结果有别于常见的抗生素生物合成基因成簇存在的现象。

1.2 链霉菌基因组岛的识别

同源重组或外源 DNA 片段插入常导致链霉菌基因组的高度不稳定性。可移动遗传元件通过水平转移整合到链霉菌染色体上, 所携带的新性状有助于宿主在特定生境下获得优势。常见的可移动遗传元件包括整合型质粒、原噬菌体、插入序列元件、转座子、整合子和染色体上一些被称为“基因组岛”(Genomic islands)的特殊区域等。有些典型岛由整合性接合元件(Integrative and conjugative elements, ICEs) 组成, 展示出位点特异性整合、切出、环化和接合转移等典型的可移动特性[6]。

外源基因组岛与宿主染色体骨架相比较, 通常在 G+C 含量、寡聚核苷酸频率、密码子使用偏好和氨基酸使用偏好等序列特征上表现异常, 这就构成了在单个基因组序列上, 计算机辅助识别岛的基础(表 1)[7]。天津大学张春霆课题组提出了一种无窗口计算 G+C 含量方法, 累积 GC 轮廓图, 通过检测 G+C 跃变区可辅助识别岛及边界[8]。随着多个链霉菌基因组完成测序, 一些基因组比较分析工具可辅助识别岛, 如 WebACT、Mauve、IslandViewer 和 MobilomeFINDER。例如, MobilomeFINDER::IdentifyIsland 工具[9]从多个基因组序列中, 分别提取出外源 DNA 插入位点上下游区域进行序列比对, 找到保守骨架从而识别出序列相似性低的区域作为外源岛。此外, PHAST 在线工具可快速识别噬菌体, 而 IS finder 工具可识别插入序列元件(表 1)。

采用 mGenomeSubtractor 工具(表 1), 我们比较分析了天蓝色链霉菌 A3(2)和其亲缘的变铅青链霉菌 TK24 的保守基因和特有基因[10]。从变铅青链霉菌 TK24 基因组注释的 7 751 个基因中, 我们识别出 472 个菌株特异性基因(*H*-value<0.81)和 9 个大岛(>10 kb)。其中, 一个 92 kb 的基因组岛 SLG 携带 φHAU3 抗性基因和 DNA 硫修饰基因簇 *dndABCDE*[11]。在该基因簇编码的 5 个 Dnd 蛋白共同作用下, DNA 骨架上磷酸二酯键中一个非桥联氧原子被硫原子取代, 形成 Rp 构象特异性磷硫酰化修饰。硫修饰的 DNA 在电泳过程中不稳定, 易发生降解, 产生一种称为 Dnd 的表型。此外, 借助 MobilomeFINDER 工具[9], 我们发现了 *dnd* 基因簇通过水平转移方式, 以基因组岛的形式广泛存在于分类地位和生态差异很大的细菌和古细菌中, 包括海洋菌、土壤菌、动植物致病菌、厌氧菌、抗菌素抗性菌等。综合分子遗传、生物化学和蛋白功能等多方面实验和生物信息学数据, 我们建立了 DNA 硫修饰的开放数据库 *dnd*DB[12], 收录了来自 29 个物种的 32 个 *dnd* 基因簇, 140 个 *dnd* 基因及蛋白的相关信息。这些分析将有助于

表 1 链霉菌基因组序列分析的常用生物信息学资源
Table 1 Bioinformatics resources used in *Streptomyces* genome analysis

工具/数据库 Tool/Database	特点，网址 Note and URL
基因组序列和注释 Genome sequences and annotation	
NCBI Microbial Genomes	NCBI 维护的细菌基因组数据库，下载基因组序列和注释 http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
The Genomes On Line Database (GOLD)	收录细菌基因组测序项目，可跟踪不同菌株测序进展 http://www.genomesonline.org
Streptomyces coelicolor	John Innes Centre 维护的模式基因组数据库 http://strepdb.streptomyces.org.uk
Streptomyces avermitilis	http://avermitilis.ls.kitasato-u.ac.jp/
Streptomyces griseus	http://streptomyces.nih.go.jp/
基因组序列比对 Genomic sequence alignment	
MUMmer	两个基因组序列超快比对，可用于 Contigs 和参照基因组比对，程序需本地运行 http://mummer.sourceforge.net
Mauve	多基因组比对时考虑了重排、插入和倒置等，可视化效果好，程序需要本地运行 http://gel.ahabs.wisc.edu/mauve
webACT	在线，直观易用，可在线 BLASTn 比对至多 5 个基因组 http://www.webact.org
mGenomeSubtractor Tools	在线，超快，常用于识别菌株保守和特有基因，可同时比较多个基因组 http://bioinfo-mml.sjtu.edu.cn/mGS
CGView Server	在线，常用于绘制多个基因组的 BLASTn 比对图 http://stothard.afns.ualberta.ca/cgview_server/
可移动遗传元件识别 Identification of movable genetic elements	
IslandViewer	在线，整合 IslandPick, IslandPath-DIMOB 和 SIGI-HMM 三种基因组岛预测工具 http://www.pathogenomics.sfu.ca/islandviewer
MobilomeFINDER	在线，识别插入 tRNA 基因等位点的基因组岛，常用于辅助 PCR 检测等实验分析 http://db-mml.sjtu.edu.cn/MobilomeFINDER
累积 GC 轮廓图 Cumulative GC contour map	在线，计算基因组 G+C 含量跃变点以识别岛的边界 http://tubic.tju.edu.cn/zcurve
δρ-web	在线，计算岛相对于全基因组的二核苷酸分布偏向性 http://deltarho.amc.nl
PHAST	在线, Prophage 快速识别 http://phast.wishartlab.com
IS Finder	插入序列 (IS elements) 数据库 http://www-is.biotoul.fr
ICEberg	整合性接合元件数据库，包括放线菌型 http://db-mml.sjtu.edu.cn/ICEberg
TADB	2 型毒素-抗毒素系统数据库，和维持元件遗传稳定性相关 http://bioinfo-mml.sjtu.edu.cn/TADB
dndDB	岛编码的 DNA 磷硫酰化系统数据库 http://db-mml.sjtu.edu.cn/dndDB

了解 *dnd* 基因簇的来源与进化, 和揭示 DNA 骨架上发现的第一种生理性修饰的生物学意义。

相似地, 我们识别了天蓝色链霉菌 A3(2)染色体上的基因组岛。例如, tRNA-Pro 基因位点插入有一个 154 kb 岛, 两端有正向重复序列, 编码一个整合酶, 并携带砷抗性基因簇(*arsOBRCT*)和与岛遗传稳定性相关的 Xre-GNAT 毒素-抗毒素基因座位(*SCO6802-SCO6803*)[13]。我们前期实验发现, 天蓝色链霉菌比变铅青链霉菌具有更高的砷耐受性。值得关注的是, 在天蓝色链霉菌中, 砷抗性 *ars* 基因簇有两个拷贝, 一个位于染色体骨架上; 另一个位于 SCO-GI14 岛上, 可能提高了宿主对砷的耐受。而变铅青链霉菌只有一个 *ars* 基因簇, 位于染色体骨架上。

1.3 放线菌型整合性接合元件的识别

整合性接合元件(ICEs)是一种广泛存在于细菌中可自主移动的典型基因组岛, 在细菌进化中具有重要意义, 但目前所发现的 ICEs 仅如冰山一角[6]。ICE 通常编码整合酶、完整的接合系统和调控蛋白等。这些蛋白可控制 ICE 从供体菌染色体上切除和环出, 以接合转移的方式进入受体菌并整合到其染色体上。我们开发的整合性接合元件数据库 ICEberg 通过文献挖掘和生物信息学预测, 系统地识别了上百个细菌中 400 多个 ICEs, 提出了 ICEs 家族分类方法(表 1)[14]。

放线菌型 ICEs 一般用 AICEs (Actinobacteria ICEs)表示。位点特异性整合发生在 ICE 的 *attB* 位点与染色体上 *attP* 位点之间。在革兰氏阴性菌和部分革兰氏阳性菌中, ICE 往往依赖四型分泌系统以单链的形式进行接合转移。而 AICE 的接合转移则是通过菌丝体之间的紧密接触完成的, 且在接合转移过程中直接以双链的形式进入受体菌。AICE 的独特之处还在于其具有自主复制功能, 且这种自主复制功能与接合转移的过程紧密相关。AICEs 大小分布在 9−24 kb, 基本都具备四个核心模块: 整合/切除、复制、接合转移和调控模块。

我们使用隐马尔科夫模型来表征与 AICE 核心模块相关特征蛋白的序列保守性[15]。对 AICE 预测结果表明, 在 12 个已测序的链霉菌中, 其中 10 个链霉菌整合了一个或多个不同类型的 AICEs, 其中含有 AICEs 最多的是天蓝色链霉菌; 而与天蓝色链霉菌近缘的变铅青链霉菌则没有发现 AICE, *Streptomyces* sp. SirexAA−E 也没有识别出 AICE。识别出的 29 个 AICEs 中, 有 26 个位于染色体上; 另外 3 个 AICEs 是质粒。比较分析发现, 染色体上的 26 个 AICEs 都位于核心区(图 1), 这一分布特征表明可自行移动的 AICEs 是链霉菌基因组多样性的重要影响因素之一。

天蓝色链霉菌 A3(2)是目前所发现含有 AICEs 最多的链霉菌。其染色体大小为 8.6 Mb, 核心区上有 4 个完整的 AICEs。其中, SLP1 是研究最详细的 AICE 之一。它的大小为 17 kb, 整合在染色体的 tRNA-Tyr 基因位点, 编码来自噬菌体的酪氨酸重组酶家族的整合酶, 同时还有切除酶辅助其从染色体环出, 接合转移依赖于其包含有 FtsK-SpoIIIE 功能域的 Tra 蛋白, 整套系统可以保证 SLP1 从天蓝色链霉菌高效率的转移到变铅青链霉菌中。此外, SPL1 涉及 DNA 磷硫酰化修饰的限制系统, 岛上的一个基因 *SCO4631* 编码了一个 IV 型核酸内切酶 ScoMcrA[16]。ScoMcrA 能够识别磷硫酰化位点, 并在距离位点 24−32 bp 的位置进行切割; 同时, ScoMcrA 还能够识别被 Dcm 甲基化的位点并进行切割, 然而对于没被修饰的 DNA 却并不切割。细菌的 DNA 硫修饰及其限制系统分别由两个不同岛编码, 令人惊讶: 一些细菌获得了 *dnd* 岛, 拥有了 DNA 磷硫酰化修饰系统(如变铅青链霉菌[11]); 同时, 另一些细菌通过基因水平转

移拥有了抵抗这些“强势修饰”的保护机制 (如天蓝色链霉菌[16])。这些有趣的现象吸引我们要系统地来研究在新环境中或各种逆境压力下, 宿主产生的应答对 DNA 磷硫酰化修饰及其限制系统的作用。

2 链霉菌次生代谢物生物合成基因簇的预测

2.1 次生代谢物及生物合成相关的常用生物信息学资源

近年来, 重要抗生素产生菌的全基因组测序(Completely sequenced)或基于基因组草图的扫描(Genome scanning)已成为快速获得抗生素生物合成基因簇的重要策略之一。在已全基因组测序的天蓝色链霉菌基因组中, 有 20 个次生代谢物生物合成基因簇; 而阿维链霉菌、灰色链霉菌和卡特利链霉菌基因组则分别有 34、30 和 38 个次生代谢物基因簇。

与链霉菌次生代谢物及其合成相关的生物信息学资源也大量涌现出来。近期建立的一些开放二级数据库为链霉菌实验人员提供了很好的研究平台(表 2)。同时, 以链霉菌为研究体系的分子遗传学、酶学、代谢工程、合成生物学等多学科需求, 也为生物信息学和化学信息学

表 2 微生物次生代谢物及生物合成分析的常用生物信息学资源
Table 2 Bioinformatics resources for natural compounds isolated from *Streptomyces* species

工具/数据库 Tool/Database	特点, 网址 Note and URL
开放数据库 Open Database	
StreptomeDB	1 900 多个链霉菌产生的 2 400 多个天然化合物, 包括活性和基因等信息, 2013 年开放 http://www.pharmaceutical-bioinformatics.de/streptomedb
DoBISCUIT	72 个已知 PKS 基因簇, 经人工校正, 2013 年开放 http://www.bio.nite.go.jp/pks
ClusterMine360	200 多个 PKS/NRPS 生物合成基因, 185 个化合物家族, 2013 年开放 http://www.clustermine360.ca
NORINE	1 000 多个 NRPS 化合物信息, 2008 年开放, 更新至 2012 年 http://bioinfo.lifl.fr/norine
PKMiner	25 个放线菌的 40 个 Type II PKS 基因簇和 231 化合物, 2012 年开放 http://pks.kaist.ac.kr/pkminer/
在线预测工具 Online prediction tool	
antiSMASH	多种次生代谢物基因簇的在线快速预测, 2011 年发表 1.0 版本, 2013 年更新至 2.0 版本 http://antismash.secondarymetabolites.org
PKS/NRPS Analysis Web-site	识别 PKS 和 NRPS 的多个功能域 http://nrps.igs.umaryland.edu/nrps
SBSPKS (or NRPS-PKS)	可预测 PKS 和 NRPS-PKS 杂合的功能域 http://www.nii.ac.in/~pksdb/sbspks/master.html
NRPSPredictor	基于保守 A 区域预测 NRPS 合成基因 http://ab.inf.uni-tuebingen.de/software/NRPSpredictor
BAGEL2	细菌素预测工具, 2013 年更新至 3.0 版本 http://bagel2.molgenrug.nl/index.php/bagel3
ThioFinder	硫肽合成基因簇预测工具, 内置已知硫肽化合物数据库, 2012 年开发 http://db-mml.sjtu.edu.cn/ThioFinder

的在线工具和二级数据库的开发提供了新的科学问题支撑。例如, StreptomeDB 数据库收录了 1 900 多个链霉菌产生的 2 400 多个天然化合物, 包括化合物、活性和基因等信息[3]。ClusterMine360 数据库系统地收录了微生物中 200 多个 PKS/NRPS 生物合成基因和 185 个化合物家族[17]。DoBISCUIT 数据库则侧重于维护微生物 PKS 基因簇的人工校正, 现版本包括了 72 个已知 PKS 基因簇[18]。此外, 多种在线预测工具为用户分析抗生素产生菌的 DNA 序列提供了方便快捷的服务(表 2)。例如, antiSMASH 工具能够快速预测多种次生代谢物及其基因簇[2], 包括 NRPS、PKS 和 Lantibiotics, 2011 年发表后论文已被引用 60 余次, 2013 年更新至 2.0 版本。BAGEL2 工具可分析多种类型的细菌素(Bacteriocins)[19]; 而我们课题组和中国科学院上海有机化学研究所刘文课题组合作开发的 ThioFinder 工具则针对其中的硫肽类抗生素(Thiopeptides)[20]。

2.2 硫肽生物合成基因簇预测工具的开发

硫肽是一类被高度修饰的、富含硫的微生物源环肽类抗生素, 对革兰氏阳性病原菌有很强抑制作用[21]。近期还发现其对人癌细胞的增殖有明显抑制作用。其成员都具有一个以六元杂环为核心、由多个五元杂环和脱水氨基酸组成的典型环肽结构。从生物合成途径上看, 硫肽类抗生素是以一条来源于核糖体的前体肽为底物, 经过一系列保守的翻译后修饰而形成特征性的框架结构。

我们收集了目前已经报道的 11 个硫肽类抗生素的基因簇, 发现了硫肽基因簇具有一些明显的分布规律。首先, 它们都具有一个或多个核糖体来源的前体肽基因(*prep* gene), 其长度不超过 120 个氨基酸、被剪切为引导肽(Leader peptides)和结构肽(Structural peptides)两部分。结构肽序列与其对应硫肽类抗生素的肽链骨架完全吻合, 富含半胱氨酸、丝氨酸或者苏氨酸(图 2B)。根据目前已知的 38 个结构肽序列, 我们使用 MEME 工具获得了描述结构肽的正则表达式“SCTT[CS][GI]CT[CS]S[CS]”。其次, 前体肽基因与一系列形成硫肽抗生素的高度保守的后修饰基因(*thiosf* gene)成簇排列。硫肽特征性框架的形成主要涉及如下三类关键酶(以 Nosiheptide 基因簇 *nosDEFGHOM* 为参考[22], 图 2A): (1) YcaO-like 蛋白(NosG-like 环化脱水酶或 NosF-like 脱氢酶), 主要负责噻唑和恶唑的形成; (2) Lantibiotic-type 脱水酶 (NosD-like 或 NosE-like 脱水酶); (3) NosO-like 或 NosH-like 蛋白, 其功能未知, 推测可能与六元含氮杂环的形成有关。我们开发的 ThioFinder 工具就是基于以上规律在 DNA 序列中识别硫肽合成基因簇(图 2B)。

ThioFinder 工具采用隐马尔科夫模型来表征硫肽合成基因簇中关键酶的序列保守性[20]。从蛋白质家族数据库 Pfam 26.0 中分别提取 NosD、NosE、NosF 和 NosG 所在蛋白家族的隐马尔科夫模型。由于 NosH、NosO 和 Prep (或 NosM 即前体肽)并没有在 Pfam 中找到所属家族, 因此我们分别提取 11 个已知基因簇中的对应基因, 使用 HMMER3.0 工具包构建了对应的隐马尔科夫模型。若一段 25 kb DNA 序列中, 同时拥有前体肽基因以及编码硫肽特征性框架的 *nosG/nosF*、*nosD/nosE* 和 *nosO/nosH* 基因, ThioFinder 就推测此段 DNA 序列含有硫肽合成基因簇。

此外, ThioFinder 还根据基因型和化学型的关联性, 以基因簇组成来为推测硫肽化学结构提供一些线索: (1) 含有 *nosL*-like 基因的 Type I 基因簇, 对应硫肽侧链中含有由 L-色氨酸衍生的 MIA 功能单元; (2) 含有 *tsrT*-like 和

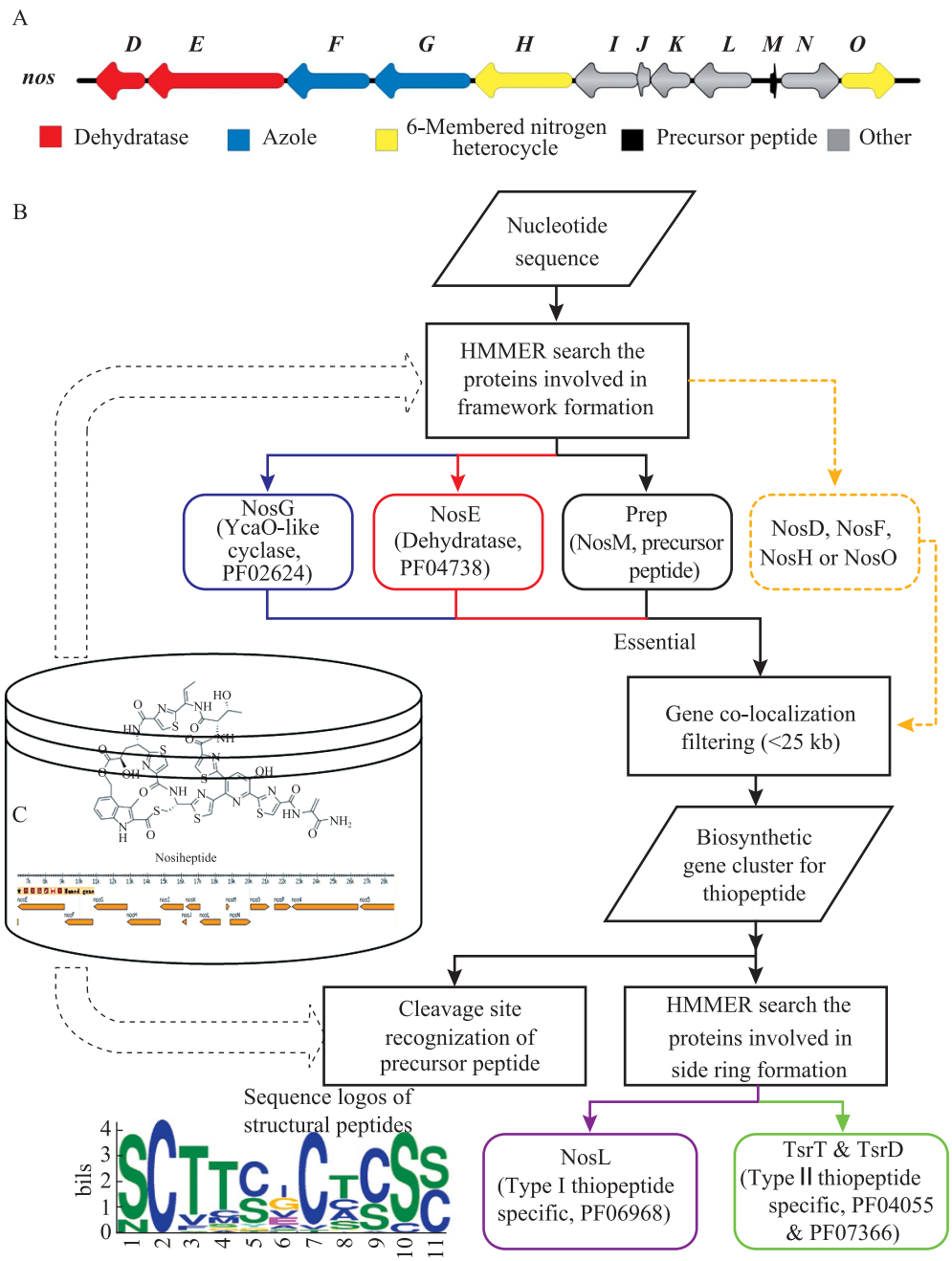

图 2　硫肽类抗生素生物合成基因簇的预测工具 ThioFinder

Fig. 2　A web-based tool ThioFinder for identification of thiopeptide gene clusters in DNA sequences

注: A: 诺西肽(Nosiheptide)的生物合成基因簇; B: ThioFinder 工具的预测方法; C: 已知硫肽化合物及基因簇的后台数据库 ThioBase.

Note: A: Organization of the biosynthetic genes, as exemplified by that for nosiheptide; B: Schematic modular pipeline of ThioFinder; C: An embedded database ThioBase containing the detailed information of thiopeptides, regarding the chemical structure, biological activity, biosynthetic gene cluster and reference.

tsrD-like 基因的 Type II 基因簇, 对应硫肽侧链中含有由 L-色氨酸衍生的 QA 功能单元; (3) Type III 基因簇不含有上述特殊基因, 化学结构中也没有由 L-色氨酸衍生的特殊功能单元。

我们在高性能四路服务器上开发了识别硫肽合成基因簇的快速工具 ThioFinder。它可在 2−3 min 内快速地识别出 *Bacillus cereus* ATCC 14 579 基因组(5.4 Mb)中完整的 Thiocillin 合成基因簇及其前体肽的剪切位点。从 NCBI 收录的 1 686 个全测序和 1 875 个部分测序的菌株基因组序列中, ThioFinder 识别出 65 个硫肽基因簇; 其中 11 个与已得到实验报道的完全一致, 其余的 54 个是首次报道。含有硫肽基因簇的菌株, 有的属于高 G+C 含量的 *Streptomyces* (链霉菌属)和低 G+C 含量的 *Bacillus* (杆状菌属)、有的是人体阴道寄生菌 *Lactobacillus gasseri* (格氏乳杆菌)和人类病原菌 *Streptococcus pneumoniae* (肺炎链球菌), 还有的是从人类粪便中分离到的 *Thermobispora bispora* (双孢热多孢菌)和从深海沉淀物中分离到的 *Verrucosispora maris* (疣孢菌)。由此可见, 硫肽产生菌分布广泛且存在于不同的生境, 可能产生新结构的硫肽类抗生素。

此外, ThioFinder 内置的基于 PostgreSQL 的开放数据库 ThioBase 首次系统地收录了 99 个已知硫肽化合物的化学结构、生物活性和产生菌等重要信息, 可供硫肽抗生素研究者对相关化合物信息进行快速查找与比较。由此, ThioFinder 提供了一个界面友好和快速高效识别硫肽合成基因簇的网上工具, 有助于更多新的硫肽类抗生素的发现和组合生物学研发新的临床药物。

3 小结

越来越多来源于不同生境的链霉菌已完成或正在进行基因组测序。生物信息学辅助的比较基因组分析可将研究重点迅速地聚焦在某株菌的可移动遗传元件和次生代谢物生成基因簇, 确定其对应的菌株特有表型和新型化合物合成机制。本文收集了近期开发的一些相关的常用生物信息学资源。以链霉菌染色体核心区和两臂的划分、天蓝色链霉菌和变铅青链霉菌基因组岛的识别、卡特利链霉菌巨型质粒为例, 简介了这些生物信息学工具和二级数据库的使用。此外, 还简述了我们课题组进行放线菌型整合性接合元件识别和开发硫肽生物合成基因簇预测工具和二级数据库的一些方法。这些生物信息学资源将有助于我们更为深入地理解链霉菌遗传和化学多样性。

参 考 文 献

[1] Hopwood DA. Soil to genomics: the *Streptomyces* chromosome[J]. Annual Review of Genetics, 2006, 40: 1−23.

[2] Medema MH, 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(Web Server issue): W339−W346.

[3] Lucas X, Senger C, Erxleben A, et al. StreptomeDB: a resource for natural compounds isolated from *Streptomyces* species. Nucleic Acids Research, 2013, 41(Database issue): D1130−D1136.

[4] Bentley SD, Chater KF, Cerdeño-Tárraga AM, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)[J]. Nature, 2002, 417(6885):141−147.

[5] Zhao C, Li P, Deng Z, et al. Insights into fluorometabolite biosynthesis in *Streptomyces cattleya* DSM46488 through knockout mutants[J]. Bioorganic Chemistry, 2012, 44(1): 1−7.

[6] Wozniak RA, Waldor MK. Integrative and conjuga-

tive elements: mosaic mobile genetic elements enabling dynamic lateral gene flow[J]. Nature Reviews Microbiology, 2010, 8(8): 552−563.

[7] Langille MG, Hsiao WW, Brinkman FS. Detecting genomic islands using bioinformatics approaches[J]. Nature Reviews Microbiology, 2010, 8(5): 373−382.

[8] Zhang CT, Zhang R, Ou HY. The Z curve database: a graphic representation of genome sequences[J]. Bioinformatics, 2003, 19(5): 593−599.

[9] Ou HY, He X, Harrison EM, et al. MobilomeFINDER: web-based tools for *in silico* and experimental discovery of bacterial genomic islands[J]. Nucleic Acids Research, 2007, 35(Web Server issue): W97−W104.

[10] Shao Y, He X, Harrison EM, et al. mGenomeSubtractor: a web-based tool for parallel *in silico* subtractive hybridization analysis of multiple bacterial genomes[J]. Nucleic Acids Research, 2010, 38(Web Server issue): W194−W200.

[11] He X, Ou HY, Yu Q, et al. Analysis of a genomic island housing genes for DNA S-modification system in *Streptomyces lividans* 66 and its counterparts in other distantly related bacteria[J]. Molecular Microbiology, 2007, 65(4): 1034−1048.

[12] Ou HY, He X, Shao Y, et al. dndDB: a database focused on phosphorothioation of the DNA backbone[J]. PLoS ONE, 2009, 4(4): e5132.

[13] Shao Y, Harrison EM, Bi D, et al. TADB: a web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea[J]. Nucleic Acids Research, 2011, 39(Database issue): D606−D611.

[14] Bi D, Xu Z, Harrison EM, et al. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria[J]. Nucleic Acids Research, 2012, 40(Database issue): D621−D626.

[15] 徐珍, 毕德玺, 李鹏, 等. 链霉菌基因组中放线菌型整合性接合元件的识别[J]. 微生物学通报, 2013, http://www.cnki.net/kcms/detail/11.1996.Q.20131011.1216.001.html.

[16] Liu G, Ou HY, Wang T, et al. Cleavage of phosphorothioated DNA and methylated DNA by the type IV restriction endonuclease ScoMcrA[J]. PLoS Genetics, 2010, 6(12): e1001253.

[17] Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis[J]. Nucleic Acids Research, 2013, 41(Database issue): D402−D407.

[18] Ichikawa N, Sasagawa M, Yamamoto M, et al. DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters[J]. Nucleic Acids Research, 2013, 41(Database issue): D408−D414.

[19] de Jong A, van Heel AJ, Kok J, et al. BAGEL2: mining for bacteriocins in genomic data[J]. Nucleic Acids Research, 2010, 38(Web Server issue): W647−D651.

[20] Li J, Qu X, He X, et al. ThioFinder: a web-based tool for the identification of thiopeptide gene clusters in DNA sequences[J]. PLoS One, 2012, 7(9): e45878.

[21] Bagley MC, Dale JW, Merritt EA, et al. Thiopeptide antibiotics[J]. Chemical Reviews, 2005, 105(2): 685−714.

[22] Yu Y, Duan L, Zhang Q, et al. Nosiheptide biosynthesis featuring a unique indole side ring formation on the characteristic thiopeptide framework[J]. ACS Chemical Biology, 2009, 4(10): 855−864.