中国农学通报 2009,25(16):280-283
Chinese Agricultural Science Bulletin

PKS–NRPS数据库的研究进展

姚 琳,郭东林
(哈尔滨师范大学,哈尔滨 150025)

摘 要:非核糖体多肽合成酶(NRPS)和聚酮合成酶(PKS)是调控次级代谢物聚酮类物质和非核糖体多肽合成的关键酶。据推测,至少有1000种聚酮/多肽化合物已经被发现,理论分析表明超过10亿的化合物能被合成。目前,国外通过高通量筛选、计算机预测-文献挖掘等方法,获得了大批量的PKS/NRPS基因及化合物结构,并由此构建和开发了一些PKS/NRPS数据库和软件,作者分析了这些数据库包括Norine、NRPS-PKS、ASMPKS和软件Cluscan、Clusean的相关信息及并概述了这些数据库及软件的应用。

关键词:PKS;NRPS;数据库;生物信息学
中图分类号:S476+8 **文献标识码:**A **论文编号:**2009-0684

Advances in PKS–NRPS Databases

Yao Lin, Guo Donglin
(*Harbin Normal University*, Herbin 150025)

Abstract: Polyketides and nonribosomal peptides are extremely important classes of bioactive secondary metabolites, and controlled by PKS and NRPS. To date, less than 10000 PKS/NRPS structures have been discovered experimentally; however, the theoretical analysis of PKS/NRPS biosynthesis performed suggests that over a billion possible structures can be synthesized. A vast number of PKS/NRPS sequences and their compound have been identified and the information was organized and hosted in many PKS/NRPS databases with the help of high through put screening technologies, computational predictions and literature–mining processes. This Review will first introduce public databases and software, including Norine、NRPS–PKS、ASMPKS、Cluscan、Clusean. Finally we will discuss the application of these databases and software。

Key words: PKS, NRPS, databases, bioinformatics

0 引言

聚酮类化合物(polyketide,PK)和非核糖体多肽(nonribosomal peptide,NRP)都是一大类由细菌、真菌和植物将低聚物通过连续的缩合反应产生的天然产物,是当前国际化学与生物学交叉学科研究的热点之一,也正在发展成为药物创新超常规的重要手段。聚酮类化合物和非核糖体多肽的生物合成涉及一列的酶促反应,组成这些合成途径的酶称为聚酮合成酶(polyketide biosynthase,PKS)和非核糖体多肽合成酶(nonribosomal peptide synthetase, NRPS)。

PKS分为三类,I类PKS也称为模块(modular)类,合成大环内酯类抗生素,如红霉素(erythromycin)、西罗莫司(sirolimus,rapamycin)、利福霉素(rifamycin)等。主要包含以下活性单元:酰基转移酶单元(acyltransferase, AT)、酰基转运蛋白单元(acyl carrier protein, ACP)、酮酯酰基合酶单元(ketoacyl synthase, KS)、酮还原酶单元(ketoacyl reductase, KR)、烯基还原酶单元(enoylreductase, ER)、脱水酶单元(dehydratase, DH)、甲基转移酶单元(methyltransferase, MT)和硫酯酶单元(thioesterase, TE);II类PKS也称为叠代(iterative)或芳香(aromatic)类,主要合成芳香族的化合物,如蒽环(anthracycline)类及四环(tetracycline)类化合物。通常

基金项目:黑龙江省教育厅"小麦单染色体微克隆体系的建立"(10541089)
第一作者简介:姚琳,女,1981年出生,黑龙江哈尔滨,助教,博士研究生,研究方向为遗传学。通信地址:150025 哈尔滨市利民经济开发区师大南路1号哈尔滨师范大学生命科学与技术学院,E-mail:yaolin19810329@163.com。
通讯作者:郭东林,女,1973年出生,吉林白城人,博士,副教授,从事遗传学研究。通信地址:150025 哈尔滨师范大学生命科学与技术学院
收稿日期:2009-04-01,**修回日期:**2009-04-16。

包括AT、ACP、KS、KR、ER、DH、MT和TE或Claisen型环化酶单元(Claisen cyclase, CYC)。不同的酶有不同的活性单元组合,但至少都包括KS、AT、ACP[1];Ⅲ型PKS又称查儿酮型PKS,和其他两种PKS不同,Ⅰ型和Ⅱ型PKS常常通过ACP活化酰基-CoA的底物,而Ⅲ型PKS直接作用于酰基-CoA活化的简单羧酸。但所有类型的PKS都是通过酰基-CoA的脱羧缩合和KS结构域或亚基催化C-C键的形成[2]。

NRPS由一系列的模块顺次排列组成,大多数NRPS的模块数为4~10个,但也有高达50个的。模块的特异结构域具有特定的酶活,催化相应单体结合到新生链肽中。这些结构域包括腺苷酰化结构域(A结构域)、巯基化结构域(T结构域)、缩合结构域(C结构域)、差向异构结构域(E结构域)和甲基化结构域(M结构域)等。微生物利用NPRS,以非核糖体途径合成天然多肽,这些多肽是一系列低分子量的次级代谢产物,具有生物活性,可被用作抗生素、免疫抑制剂、抗癌和抗病毒因子、铁载体及生物表面活性剂等。多肽的装配由NRPS的多模块酶以硫模板机制完成[3]。目前发现的NRPSs可分为3类:A型——线性NRPS;B型——重复型NRPS;C型——非线性NRPS[4]。

PKS/NRPS杂合抗生素是由PKS和NRPS共同参与完成的。Ⅰ型PKS和NRPS具有相似的结构和功能特点,它们都是具有模块结构的多功能的巨型酶,这种结构、催化机制惊人的相似性促使人们寻找整合了NRPS和PKS模块的杂交NRPS2PKS体系。既然NRPS和PKS的模块式构造已被成功地运用在组合生物合成中合成多种天然、非天然的产物,可以想象能将氨基酸、短的羧酸整合为最终产物的杂交NRPS/PKS体系能够产生更多的化学结构多样性[5],如雷帕霉素、埃坡霉素和博莱霉素等[6]。

由于这些抗生素在抗感染、抗肿瘤及免疫抑制等方面的巨大应用价值,吸引了许多科学工作者对PKS和NRPS做深入研究。目前,人们集中于在分子水平上利用基因工程方法改造NRPS和PKS,如结构域替代、添加和突变,以期产生具有新结构和功能的化合物,满足医药上的需求[7-10]。

利用生物信息学方法对NRPSs/PKSs序列分析预测是设计新的非核糖体多肽/聚酮化合物的必要手段[11-15]。随着NRPS和PKS研究的不断深入,科研工作者建立了NRPS和PKS数据库,我们可以更加方便地对NRPS和PKS进行各种研究,包括序列比对、分析,结构预测,结构和功能关系的研究,分子改造设计等。

1 Norine(NOnRibosomal peptides)数据库

这是首个完整的NRP数据库。由法国巴黎第十一大学生命信息研究组和麻省理工学院以及法国里尔的ProBioGEM lab的NRPS梯队维护。网址为http://bioinfo.lifl.fr/norine/index.jsp。

Norine数据库搜集了七百多种NRP,覆盖了所有的结构及活性类型,并且种类还在增加。

Norine提供的研究NRP系统的计算工具,可用于获得NRPS的各种代谢产物及其结构,为新药的设计奠定理论基础。

Norine的主页上有5个按纽,分别是Home、DataBase、Monomers、Help、Surnit/Modify。Home是对Norine的简介。DataBase中存储NRP数据,通过search可以方便检索到NRP。检索分为一般检索和结构检索。一般检索条件包括多肽名称、生物活性、结构类型、多肽分子量、引文检索和物种检索,输入一种或几种检索条件,可以得到多肽列表。结构检索基于多肽的结构特征比对多肽单体,分为多肽组合检索和多肽结构检索,通过检索可得到一种多肽的所有单体和多肽的结构模式图(或线性示意图)。使用Norine数据库可方便获得所有已知NRPs的各种信息,此数据库数据为定期更新[16]。

2 NRPS–PKS数据库

NRPS-PKS数据库由印度国立免疫学研究所维护。网址是http://www.nii.res.in/nrps-pks.html。主要用于分析NRPS,PKS和NRPS/PKS杂合基因簇及其结构特征。

此数据库是由Mohd、Zeeshan、Ansari等人将其已建立的NRPS-SEARCHNRPS数据库和PKS-SEARCHPKS[17]数据库结合所构建。搜集了307个ORFs,2223个功能蛋白结构域,68个起始/延伸前体和四类家族101个化合物的化学结构。NRPS-PKS数据库分成四组独立子数据库,NRPSDB、PKSDB、ITERDB和CHSDB。这四组子数据库根据NRPs,PKs和NRPs/PKs化合物结构域种类,结构域和其连接序列,起始和延伸结构域种类以及相应结构域活性位点残基所划分的。

点击NRPS-PKS主页,有五个界面链接。点击ANALYZE HYBRID NRPS/PKS SEQUENCES可进入化合物快速分析界面,通过序列比对分析未知氨基酸序列。点击NRPSDB、PKSDB、ITERDB和CHSDB链接,分别进入相应子数据库界面[18]。

2.1 NRPSDB

NRPSDB数据库搜集了17个NRP簇和5个杂合

簇。进入主界面后，左面是是22个化合物的链接，使用者可找到任意一种化合物的化学结构、NPRS的开放读码框，A结构域的特异性底物，以及结构域构造图、链接区序列、A和C结构域活性位点残基。同时，左边的搜索键可对假定蛋白序列与22个化合物的蛋白序列进行比对，以期预测假定蛋白的结构域、基因序列和连接区序列。

2.2 PKSDB

此数据库搜集了19种模块型PKS序列。有三个主要的功能，预测功能域、AT功能域底物特异性分析、功能域鉴定和底物预测的定位[19]。

与NCBI的CDD数据库[20]相比，PKSDB数据库是为预测PKS功能域而专门设计的。PKSDB拥有能检测多种蛋白结构域的特殊位点记分矩阵（PSSMs），可正确的检测所有的还原功能域。

PKSDB的AT功能域底物特异性分析功能PKSDB能通过鉴定AT功能域的残基可鉴定AT功能域的残基，并预测其特殊的起始和延伸单元。它的分子模型计算功能不仅解释了AT功能域的底物立体选择的分子基础，并能为选择多种底物的AT功能域的活性位点残基提供立体化学原理。用SCWRL程序做霉菌素的PKSs簇AT功能域分子模型预测，发现第200位残基在选择丙二酸和甲基丙二酸作为底物时，甲基碳原子与200位残基上Phe环的$C^{\delta 2}$和$C^{\varepsilon 2}$碳原子发生空间碰撞，Phe200与甲基丙二酸空间不兼容，因此含有Phe200的AT功能域只能选择丙二酸作为底物。相反，选择甲酰基丙二酰辅酶A做底物的AT功能域是由于较大的Phe200变成相对小的Ser残基，避免了与甲基碳的空间碰撞。此外，针对选择多种特别的底物，大多数的AT结构域的第200位点上有一些小的氨基酸残基如Gly、Ser和Ala，相应的底物比甲基丙二酸大。

PKSDB也可用于鉴定多种微生物基因组中的PKS簇，并能预测其功能域结构及底物种类。

2.3 ITERDB

用于分析预测地迭代型PKS序列。这个数据库与PKSDB数据库结合使用。

2.4 CHSDB

CHSDB数据库用于分析预测查尔酮型PKS序列。此数据库搜集11种植物PKS蛋白，三种细菌PKS蛋白和一种未知源PKS蛋白的序列、起始延伸单元的化学结构和缩合循环的数量。

2.5 SEARCHGTR

此数据库用于分析与酰基载体蛋白特异性结合的糖基转移酶序列和预测其特异性底物[21]。这个数据库与GenBank[22]，Swiss-Prot[23]，CAZY[24]共享数据。

3 ASMPKS

ASMPKS(Analysis System for Modular Polyketide Synthesis)数据库是模块型PKS分析系统，由南韩维护，以PKSDB数据为基础，定期更新界面[25]。网址是http://gate.smallsoft.co.kr:8008/~hstae/asmpks。

此数据利用blast和ClustalW程序比对氨基酸序列，找出已知PKS簇，或通过结构域分析预测到新PKS候选基因。ASMPKS系统分为两个部分。第一部分是公共数据库，由PKS相关基因，结构域和模型构成，用于搜索和比对蛋白序列。第二部分由微生物基因组和聚酮物相关数据构成，主要用于描述特定微生物基因组中聚酮类物质的合成过程，如利用计算机模仿新模件合成，加工链的延伸和添加碳链。

目前，ASMPKS只是分析模件型PKSs的系统，不能分析PKS-NRPS型和迭代型PKSs。

4 Cluscan

Cluscan(Cluster Scanner)是分析编码NRPS、PKS和PKS/NRPS杂合酶基因簇半自动化DNA序列分析软件。也可用于预测PKS、NRPS反应产物及其结构域的化学结构，并以图形形式输出[26]。网址是http://bioserv.pbf.hr/。

Cluscan与GeneMark、HMMER和Pfam共享分析程序。目前，此软件只有30天免费试用权限。

5 Clusean

Clusean (CLUster SEquence ANalyzer)是细菌次级代谢物合成基因簇的计算机分析软件，由信息、图形用户界面构成，能与其他生物化学软件结合预测分析PKS/NRPS结构[27]。此软件由德国Tübingen大学微生物研究所Tilmann. Weber等人研发的公共数据库。用户可在http://www.mikrobio.uni-tuebingen.de/ag wohlleben/research groups/ag weber/downloads.html下载。

Clusean主要应用于蛋白相似性/同源性鉴定（NCBI nr (blastp)）、保守区的鉴定（Pfam (HMMer)），PKS/NRPS杂合基因簇megaenzymes结构域及其序列保守区的鉴定（ABDomains (HMMer)），NRPS缩合结构域分类和NRPS A结构域特异性预测（NRPS predictor）以及假定基因功能预测（ABDomains (HMMer)）。数据输入格式基于EMBL格式，分析结果易于被通用序列分析程序如Artemis接受。此软件的兼容系统有LINUX, UNIX和MS Windows系统。

6 展望

自然界中的不同的聚酮物质及非核糖体多肽在结构和生理活性上千差万别，但他们分别都是通过同

一代谢途径产生,因此聚酮类物质和非核糖体多肽的特殊合成代谢途径目前备受人们瞩目。如何在自然界中分离纯化新的化合物或是人工合成新的化合物,其分离或合成技术是研究的关键。从自然界中直接发现新的化合物由于受多种检测手段的局限而无法快速实现;因此通过分子生物学手段,利用基因工程技术改造聚酮合成酶和非核糖体多肽合成酶进而产生新的化合物则是当前人们关注的焦点。但无论是哪一种方式,都需要对化合物结构特征及代谢调控机制有着相应的了解。应用生物信息学对现有的化合物、合成酶及合成途径的分类则为新化合物的产生奠定理论基础。因此,当前如何建立海量的数据库,整合现有资源并开发新的结构域分析程序,成为人们迫在眉睫的任务。

参考文献

[1] 孙宇辉,邓子新.聚酮化合物及其组合生物合成.中国抗生素杂志,2006,31(1): 6-14+18.

[2] Funa N, Ohnishi Y, Fujii I, et al. A new pathway for polyketide synthesis in microorganisms.Nature,1999,400: 897-899.

[3] Kham N H, Roets E, Hoogmartens J, et al. Quantitative analysis of oxytetracycline and related substances by high-performance liquid chromatography. J. Chrom atogr,1987,405: 229-45.

[4] Mootz H D, Schwarzer D, Marahiel M A. Ways of Assembling Complex Natural Products on Modular Nonribosomal Peptide Synthetases.ChemBioChem,2002,3:490-504.

[5] 汪星明,赵凤生.组合生物合成.国外医药抗生素分册,2003,3,24(2), 49-53.

[6] 连云阳,程元荣.杂合NRPS-PKS的研究进展.//:中国药学会抗生素专业委员会.创新药物及新品种研究、开发学术研讨会论文集.烟台, 2006:19-25.

[7] Menzella H G, Reid R, Carney J R, et al. Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. Nature Biotechnol, 2005,23:1171-1176.

[8] Doekel S, Marahiel M A. Dipeptide formation on engineered hybrid peptide synthetases. Chem Biol,2000,7(6): 373-384.

[9] Gokhale R S, Tsuji S Y, Cane D E, et al. Dissecting and exploiting intermodular communication in polyketide synthases. Science, 1999,284:482-485.

[10] McDaniel R, Thamchaipenet A, Gustafsson C, et al. Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. Proc Natl Acad,1999,96:1846-1851.

[11] Gonzalez-Lergier J, Broadbelt L, Hatzimanikatis V. Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways. J. Am. Chem. Soc,2005,127: 9930-9938.

[12] Kamra P, Gokhale R S, Mohanty D. SEARCHGTr: a program for analysis of glycosyltransferases involved in glycosylation of secondary metabolites. Nucl.Acids Res, 2005,33:220-225.

[13] Lautru S, Deeth R J, Bailey L M, et al. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nature Chem.Biol. 2005,1(5): 265-269.

[14] Rausch C, Weber T, Kohlbacher O, et al. Specificity prediction of adenylation domains in nonribosomal peptide synthetases(NRPS) using transductive support vector machines (TSVMs). Nucl. Acids Res,2005,33:5799-5808.

[15] Yadav G, Gokhale R S and Mohanty D. Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J. Mol. Biol,2003,328:335-363.

[16] Norine Caboche S, Pupin M, et al. NORINE: a database of nonribosomal peptides. Nucleic Acids Res,2008,36:D326-D331.

[17] Ansari M Z, Yadav G, Gokhale R S, et al. NRPS-PKS: a knowledge based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res,2004,32:405-413.

[18] Yadav G, Gokhale R S, Mohanty D. SEARCHPKS: a program for detection and analysis of polyketide synthase domains. Nucleic Acids Res,2003,31:3654-3658.

[19] Yadav G, Gokhale R S, Mohanty D. Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J. Mol. Biol,2003,328:335-363.

[20] Marchler Bauer A,Panchenko A R,Shoemaker B,et al. CDD:a database of conserved domainalignments with links to domain three-dimensionalstructure. Nucl. Acids Res,2002,30:281-283.

[21] Kamra P, Gokhale R S, Mohanty D. SEARCHGTr: a program for analysis of glycosyltransferases involved in glycosylation of secondarymetabolites. Nucleic Acids Res,2005,33:W220-W225.

[22] Benson D A, Karsch-Mizrachi I, Lipman D J, et al. GenBank. Nucleic Acids Res, 2003,31:23-27.

[23] Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res,2003;31:365-370.

[24] Coutinho P M, Deleury E, Davies G J, et al. An evolving hierarchical family classification for glycosyltransferases. J.Moi.Biol, 2003, 328:307-317.

[25] Tae H, Kong E B, Park K. ASMPKS: an analysis systemfor modular polyketide synthases.BMC Bioinform,2007,8:327.

[26] Starcevic A, Zucko J,S imunkovic 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. Nucleic Acids Res, 2008,36:6882-6892.

[27] Weber T, Rausch C, Lopeza P, et al. CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J. Biotechnol, 2009,140:13-17.