Research Paper 研究报告

微生物学报 *Acta Microbiologica Sinica*
52(4):457 – 465; 4 April 2012
ISSN 0001 – 6209; CN 11 – 1995/Q
http://journals.im.ac.cn/actamicrocn

基于文献挖掘的巨大芽胞杆菌代谢网络模型的构建与分析

周冒达[1,2,3],邹伟[1,2,3],刘立明[1,2,3*],陈坚[1,2,3]

江南大学,[1] 食品科学与技术国家重点实验室,[2] 工业生物技术教育部重点实验室,[3] 糖化学与生物技术教育部重点实验室,无锡 214122

摘要:【目的】通过挖掘实验性文献,建立巨大芽胞杆菌事实型代谢网络模型,以详尽解析生理特性,优化其生理功能。**【方法】**从 PubMed、Derwent Innovations Index、中国知网等公共文献(专利)数据库中获取与巨大芽胞杆菌(*Bacillus megaterium*)相关的实验性文献建立本地文献数据库。采用文献挖掘工具获取功能基因、酶、代谢物和生化反应等信息,以其为基础构建代谢网络粗模型,进一步借助 KEGG 等数据库修正以及 Matlab 程序的模拟得到精细模型(系统生物学标记语言的形式)。**【结果】**最终的精细模型共有 292 个生化反应、378 个代谢物、220 个酶和 217 个基因。以 1.62 mmol/g cell/h 的葡萄糖底物吸收速率为限制性条件,模拟的菌体比生长速率为 0.089 h^{-1},略低于实验值 0.11 h^{-1}。此外,嘧啶代谢途径的单基因敲除模拟结果表明,准确率为 90%。**【结论】**该代谢网络模型涵盖了中心代谢途径、维生素 B_{12} 合成途径和氨基酸代谢途径,并在一定程度上反映了营养底物与基因对巨大芽胞杆菌生长性能的影响。

关键词: 巨大芽胞杆菌,文献挖掘,代谢网络模型

中图分类号: Q93 **文献标识码:** A **文章编号:** 0001-6209(2012)04-0457-09

作为一种重要的革兰氏阳性菌,巨大芽胞杆菌(*Bacillus megaterium*)广泛应用于环境、食品、医药等工业领域。该菌具有:(1)宽广的底物利用范围:不仅可利用葡萄糖、醋酸、富马酸等为碳源和能源[1],还可通过降解苯乙烯[2]、二氯苯[3]等有毒害物质而获取碳骨架和能量;(2)高效的异源蛋白表达和生产能力[4]:具有完备的蛋白质合成、修饰系统,不产生碱性蛋白酶和内毒素,维持重组质粒良好的稳定性;(3)理想的维生素生产能力:不仅自身能合成维生素 B_{12}[5],还作为 *Ketogulonicigenium vulgare* 的伴生菌用于合成维生素 C 前体 2-酮基-L-古龙酸[6]等优势。

目前,国内外围绕 *B. megaterium* 蛋白质合成的调节机制解析[7]、氨基酸的合成与分解途径[8]、营养物质对蛋白质合成与分泌的影响等方面开展了大量系统而卓有成效的研究工作。然而,上述研究只是从某一侧面解析了 *B. megaterium* 的生理特性,缺乏对其生理功能的全局阐释。基于基因组序列和比较基因组学的基因组规模代谢网络模型(Genome-scale metabolic model,GSMM)是现在

基金项目:全国优秀博士论文作者支持基金(200962);教育部新世纪优秀人才支持计划资助(NCET-10-0456);江苏省产学研前瞻性项目资助(BY2009112);国家自然科学基金重点项目(20836003);江苏省优势学科项目;111 工程(111-2-06);江苏省"六大人才高峰"高层次人才项目(2011-NY033)

* 通信作者。Tel/Fax: +86-510-85197875; E-mail: mingll@jiangnan.edu.cn

作者简介:周冒达(1987 –),男,山东临沂人,硕士研究生,从事工业微生物系统生物学研究。E-mail: chdzmd@163.com

收稿日期:2011-09-12;修回日期:2012-01-29

比较常用的全局阐释特定微生物生理功能的平台。然而,在构建巨大芽胞杆菌 GSMM 过程中发现基因组电子注释无法辨识巨大芽胞杆菌特有的一些生理特性,导致所构建的 GSMM 出现假阳性或假阴性等结果,无法真实反映 *B. megaterium* 的生理功能。另一方面,以 PubMed 为代表的文献或专利数据库中积聚了大量与 *B. megaterium* 相关的实验性论文和专利,可挖掘其中的生化信息以构建反映 *B. megaterium* 真实生理特性的 GSMM[9]。因此,本文在搭建本地实验性文献数据库的基础上,通过文献挖掘获取与 *B. megaterium* 相关的基因、酶、生化反应等信息,借助自行编写的基于 Matlab 的网络模型构建工具(参见软件著作权 2011R11L050355),构建 *B. megaterium* 代谢网络模型,以透彻理解其生理特性,为调控其生理功能奠定坚实基础。

1 材料和方法

基于文献挖掘构建 *B. megaterium* 代谢网络的方法和策略如图 1 所示。在这一过程中,从 Web of Science、PubMed 等文献或专利数据库中提取与 *B. megaterium* 相关的实验性文献,建立本地数据库。在此基础上,采用文献挖掘工具,获取与 *B. megaterium* 代谢相关的基因、酶、代谢物、转运反应及其之间的相互关系等实验性数据并进行整合,利用自行编制的 Matlab 程序,构建 *B. megaterium* 代谢网络模型,结合具体实验数据验证模型的准确性。

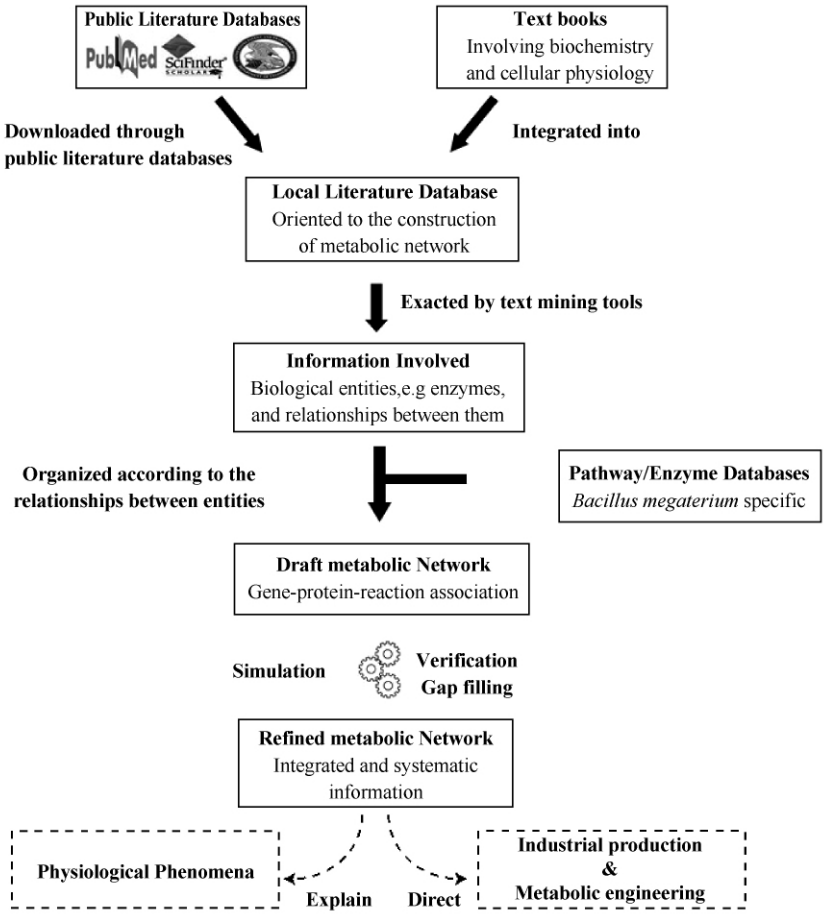

图 1 *B. megaterium* 代谢网络构建流程

Fig. 1 Workflow for the re-construction of the metabolic network of *Bacillus megaterium*.

1.1 建立本地文献数据库

以巨大芽胞杆菌、*Bacillus megaterium* 或 *B. megaterium* 为关键词从 Web of Science、PubMed、SciFinder、USPTO Patent Databases、Derwent Innovations Index、万方数据库、中国知网等公共文献(专利)数据库中下载与 *B. megaterium* 相关的实验

性文献，建立本地文献数据库。同时，与生物化学、细胞生物学或微生物生理学等相关的教科书也构成了本地数据库重要的数据来源，主要为修正所挖掘的信息提供参考。

1.2 挖掘信息

构建 *B. megaterium* 代谢网络的信息主要包括：(1)基因；(2)催化生化反应的蛋白质-酶；(3)中间代谢产物：包括区分同一代谢物的不同名称；(4)生化反应或跨膜转运反应，并确定反应方向是否可逆以及反应所属的代谢亚系统；(5)基因、酶、生化反应三者之间的对应关系；(6)亚细胞定位：分为细胞内、细胞膜、细胞外，并对中间代谢产物或酶进行亚细胞定位；(7)生物细胞组分：包括 DNA、RNA、蛋白质、细胞壁、脂质等；(8)能量维持：生长所需的 ATP 和维持基本生命活动的 ATP。

进行文献挖掘的途径主要有两类：(1)利用文献挖掘工具（表 1）从本地文献数据库中进行挖掘；(2)在相关代谢途径或酶数据库（表 2）中进行挖掘。将通过上述两种途径挖掘到的信息进行整合、借鉴。其中，数据库 Brenda、MetaCyc 中的信息均有实验性文献支持，而 Uniprot 中仅部分信息得到实验性文献支持。各挖掘工具的侧重点也有所不同，如 iHOP 主要用于获取基因信息。因此，信息挖掘顺序是：Brenda-MetaCyc-Uniprot-各文献挖掘工具。此外，对新旧数据、同一问题的表述方式都有统一处理方式。KEGG 数据库中与 *B. megaterium* 相关信息没有获得实验性文献支持，但可以 KEGG 的生化反应为基础，对其他来源获得的反应进行编号、确定反应方向、添加相应的 *B. megaterium* 基因信息、统一反应物名称以及设定生化反应所归属的代谢途径。

表 1 文献挖掘工具

Table 1 Literature mining tools used in the study

Literature mining tools	URL	Description
GoPubMed	http://gopubmed.org	Exploring PubMed with the gene ontology
MEDIE	http://www-tsujii.is.s.u-tokyo.ac.jp/medie/	Retrieving relational concepts from huge texts
Whatizit	http://www.ebi.ac.uk/webservices/whatizit/info.jsf	A suite of modules that analyses text for contained information
iHOP	http://www.ihop-net.org/UniPub/iHOP/	Providing a gene-guide network as a natural way of accessing PubMed abstracts
EBIMed	http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp	Gathering facts for proteins

表 2 代谢或酶数据库

Table 2 Pathway/Enzyme databases used in the study

Databases	URL	Description
KEGG	http://www.kegg.com/	Integrating genomic, chemical and systematic functional information
Brenda	http://www.brenda-enzymes.org	A database with detailed introduction of enzymes
MetaCyc	http://metacyc.org/	Involving different pathways
Uniprot	http://www.uniprot.org/	A universal protein database

1.3 建立可信度分数体系

将获取的生化反应信息按照方程式的形式整理于 excel 文件中，同时附加说明催化该反应的酶、基因、亚细胞位置、亚代谢系统等信息。由于相关信息的来源或查找方法不同，信息的可信度也有差别。为了能够体现这种可信度，建立一个如表 3 所示的可信度分数体系对所收集的信息进行打分评价。分数越低，则表示可信度越低。

表 3 可信度分数体系

Table 3 Confidence scoring system for assessing the mined information

Confidence score	Supporting resources
1	Needed for physiological activity or gap filling
2	Organized from vague, indirect or partial literature evidences
3	Directly exacted from databases and literatures

1.4 *B. megaterium* 网络模型的构建、精细化与模拟

通常，一种基因独立编码一种酶，而有些基因只能编码酶的亚基，所以包含多个亚基的复合酶需要由多个基因共同编码而成。此外，不同的酶（同工酶）会催化同一个反应，一种酶也会催化多种反应。所以，确定基因-酶-反应的关系时，根据催化反应的酶以及编码酶或亚基的基因，确定与反应对应的基因之间的布尔逻辑关系[10]。

将上述挖掘的信息按照基因-蛋白质-生化反应的次序建立关联，形成粗模型。粗模型中一般会有两种代谢物[11]：(1)并未通过细胞的摄取，又不能通过胞内代谢反应而生成的代谢物；(2)并未输出到细胞外，又不能通过胞内代谢活动而消耗的代谢物。

这两种代谢物很可能是代谢网络模型中的漏洞(gaps)。通过 KEGG Mapper (http://www.kegg.com/kegg/mapper.html),在不同的代谢亚系统与整个细胞的代谢网络图中标注出粗模型中的代谢反应,可以直观地找出上述两种代谢物,从而确定网络模型中的漏洞。参照 KEGG 数据库中 *B. megaterium* QM B1551 所具有的各个代谢途径,添加必要的代谢反应、酶及其编码基因(信息的可信度设为 1),填补代谢漏洞。另外,通过文献挖掘的信息确定 *B. megaterium* 的细胞生物量、用于细胞生长以及维持基本生命活动所需的 ATP,进而确定合成 *B. megaterium* 细胞组分的方程式。

采用 Excel 形式存储 *B. megaterium* 的精细代谢网络模型,参考 MetNetMaker 并采用自行编写的基于 Matlab 程序将 Excel 形式的代谢网络模型转换为系统生物学标记语言(systems Biology Markup Language,SBML)格式的文件。以葡萄糖为唯一碳源,细胞生物量合成反应为目标方程式,通过线性规划方法(linear programming solver)进行优化计算。模型中的可逆反应与交换反应的限制条件一般设为 -1000 到 1000 mmol/g cell/h,不可逆反应的设为 0 到 1000 mmol/g cell/h,氧气运输的限制条件为 -1000 到 0 mmol/g cell/h。交换反应限制条件中的负值表示物质进入细胞,正值表示物质离开细胞,零值表示无物质流。基于 Mosek 软件按照代谢流平衡分析法则[12](Flux Balance Analysis, FBA)对代谢网络模型进行模拟与修正,将修正后的精确代谢网络模型经过 Cytoscape 的读取、处理,将 SBML 文件转换成网络形式,并进行可视化。最后,以 Adobe Illustrator 形式展现所构建的代谢网络。

2 结果和分析

2.1 本地文献数据库的构建与信息挖掘

以巨大芽胞杆菌、*B. megaterium* 等为检索词从公共文献(专利)数据库中批量下载与 *B. megaterium* 相关实验性文献约 4300 篇,构建本地数据库。通过文献挖掘工具,提取获得了如表 4 所示的 203 个代谢相关基因、201 个具有 EC 号的酶、354 个代谢物和 241 个生化反应(具体信息见附件)。

2.2 基于文献挖掘的 *B. megaterium* 代谢网络粗略模型的构建

以所挖掘的信息为基础,采用自行编写的 Matlab 程序(软件著作权 2011R11L050355),建立基因-酶-生化反应三者之间的关联而构建代谢网络粗模型(以 excel 文件形式存储)。在所挖掘到的 241 个生化反应中,有 148 个生化反应(占总反应数的 61%)能与所挖掘的基因和酶完全匹配;86 个生化反应能匹配到所催化的酶,但所对应基因;7 个生化反应(占总反应数的 3%)难以匹配到酶和基因。这一结果表明,基于文献挖掘的基因-酶-生化反应具有很高的关联性。

表 4 基于本地数据库所挖掘的相关文献信息统计数据

Table 4 The statistic data origin from literature mining

Content	Number
Genes	203
From databases	186
From literatures	17
Enzymes	201
Cytoplasmic	193
Membranal	7
Extracellular	1
Metabolites	354
Cytoplasmic	344
Extracellular	10
Reactions	241
Biochemical reactions	235
Membrane transport	6
Enzymatic	3
Nonenzymatic	3

在所构建的粗模型中,约有 50 个生化反应的酶由多个基因编码,113 个生化反应所需的酶仅由单个基因编码。将上述基因进行基因本体(Gene Ontology,GO)注释,发现基因的代谢功能主要集中于中心代谢途径、脂肪酸代谢、维生素 B_{12} 合成和氨基酸代谢。

2.3 精细代谢网络模型的构建与解析

为了寻找和填补粗略代谢模型中的漏洞(gaps),作者借助 KEGG Mapper 对已构建的代谢网络粗模型进行可视化和系统分析。如在维生素 B_{12} 合成途径中,关键中间代谢产物腺苷钴啉醇酰胺不能以腺苷钴啉胺酸和氨丙醇为前体进行合成。但在 KEGG 数据库中,*B. megaterium* QM B1551 有催化这一反应的酶(磷酸化腺苷钴啉醇酰胺合成酶),其编码基因为 *cbiB*(BMQ_1995)。将这一生化反应所需的酶、EC 号(6.3.1.10)、编码基因、代谢物等信息整合进粗代谢模型中,即可填补模型中的代谢漏洞。类似地,通过对 KEGG、Brenda、本地数据库等进行

第二轮文献挖掘（表5），获得33个相关信息填补代谢漏洞。此时，所构建的代谢网络中包含274个生化反应（表5）、220个酶、217个编码基因和364个代谢物。

表5　各信息来源挖掘到的反应数

Table 5　The biochemical reactions from different literature mining sources

Sources	Quantity of reactions	Proportion/%
Literatures	180	66
Pathway/Enzyme Databases	162	60
Brenda	107	39
Uniprot	34	12
MetaCyc	64	23
Complementary	33	12
Total	274	100

在基因组规模代谢网络模型中，描述 *B. megaterium* 细胞组分的蛋白质、脂质、RNA、DNA 等数据借用枯草芽胞杆菌数据[13]。用于生长所需的ATP、用以维持基本生命活动的ATP取自稀释率为 0.11 h^{-1} 的恒化培养实验[14]，分别为 56.82 mmol ATP/g cell/h 与 3.96 mmol ATP/g cell/h。

此外，有文献[15]表明 *B. megaterium* 确实存在某一代谢物（如组氨酸）并直接参与了细胞组分合成，但未描述相应的合成途径。对于这一情况，可通过添加18个 sink reactions 实现这些代谢物的合成[10]。至此，所构建的模型中含有292个生化反应和378个代谢物。将添加了 sink reactions 的代谢网络模型借助基于 Mosek 的 FBA 模拟 D = 0.11 h^{-1} 条件下以 −1.62 mmol/g cell/h 的速率吸收葡萄糖时的比生长速率，结果为 0.089 h^{-1}，这一数据尽管与文献[14]的实际测定值 0.11 h^{-1} 低 19%，但在一定程度上描述了 *B. megaterium* 的真实生理特性。出现这一状况的原因在于该模型是基于文献挖掘的代谢网络模型，大约 80% 的生化反应来自文献挖掘，而不是通过基因组序列注释的，不涵盖全部生化反应。类似地，所构建的代谢网络模型只涉及63个代谢亚系统，这一数据小于基于 *B. megaterium* 基因组注释的 KEGG 代谢亚系统（82个）。

如图2所示，在所构建的网络模型中，催化96%生化反应（264个）的酶可分为6大类，分别是氧化还原酶（EC1）、转移酶（EC2）、水解酶（EC3）、裂合酶（EC4）、异构酶（EC5）和连接酶（EC6）。其中氧化还原酶、转移酶所催化的生化反应最多，约占总反应 28%（76）和 29%（80）；其次是裂合酶（34）、水解酶（34）、连接酶（22）、异构酶（17）。构建的网络模型所涉及的代谢物，约有 86% 源于文献挖掘，这些代谢物与中心代谢途径、氨基酸代谢、脂肪酸代谢、脂质代谢和维生素 B_{12} 合成等相关。

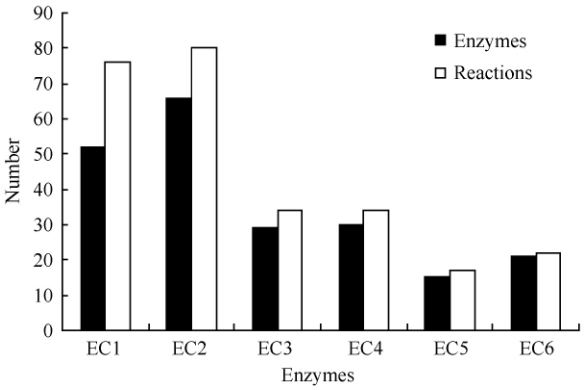

图2　酶的分类及其所催化的反应

Fig. 2　The classes of enzymes and the corresponding catalytic reactions.

2.4　单基因敲除实验

借助 FBA 计算方法并以基本培养基[14]为模拟条件，发现单独敲除47个基因后 *B. megaterium* 无法生长，而单独删除其他170个基因并不影响细胞的生长（详细基因列表见附件）。影响细胞生长的基因主要集中于中心代谢途径、核酸代谢、氨基酸合成。为此，参照 Lach DA *et al.* 的研究结果[16]，在模型中分别添加转运乳清酸（orate，OROA）、二氢乳清酸（dihydroorotate，DOROA）、乳清酸-5-磷酸（orotidine 5′-phosphate，OMP）和尿嘧啶核苷酸（uridine monophosphate，UMP）的反应，然后分别敲除如表6所示的4个基因进行模拟。这4个基因所编码的酶催化细胞内的嘧啶代谢，与上述物质的生成相关。表6的比对结果表明单基因敲除模拟准确度达到90%。

2.5　网络模型可视化分析

将获得的精细代谢网络模型的 SBML 文件采用 Cytoscape 软件进行读取和处理，得到网状结构的 *B. megaterium* 代谢网络模型，实现了代谢网络模型的可视化，并采用 Illustrator 的形式直观、简洁地展现所构建的代谢网络。直观的 Illustrator 形式代谢网络包括碳水化合物代谢反应（图3A），此外还有

表 6 实验结果与单基因敲除模拟的比对

Table 6 Comparison of computational simulation on gene deletion with experimental data

Gene	Enzyme	EC	MM	OROA	DOROA	OMP	UMP
BMQ_4254(*pyrC*)	Dihydroorotase	3.5.2.3	- / -	+ / +	+ / +	- / +	+ / +
BMQ_4255(*pyrB*)	Carbamylaspartotranskinase	2.1.3.2	- / -	+ / +	+ / +	+ / +	+ / +
BMQ_4250(*pyrD*)	Dihydroorotate oxidase	1.3.98.1	- / -	+ / +	- / -	- / +	+ / +
BMQ_4249(*pyrF*)	Orotidine-5′-phosphate decarboxylase	4.1.1.23	- / -	- / -	- / -	- / -	+ / +

experimental result / *in silico* result; -, no growth; +, growth; MM, minimal medium

氨基酸代谢、维生素 B_{12} 的合成和脂质代谢（图 3B）等重要的代谢活动。

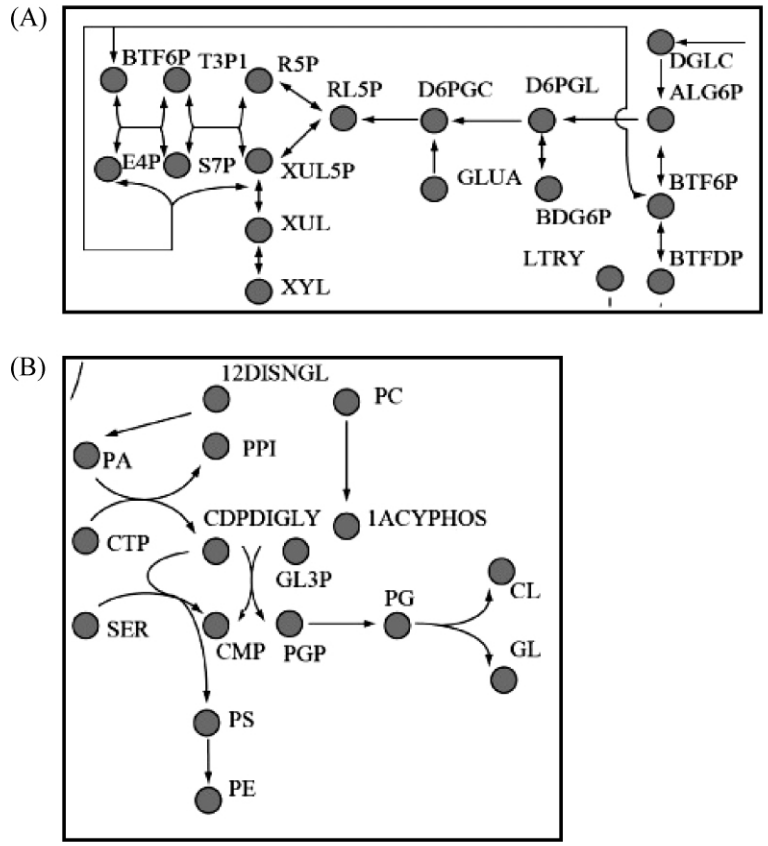

图 3 *B. megaterium* 代谢网络可视图

Fig. 3 The metabolic network in the cytosol of *Bacillus megaterium*. A, Pentose phosphate pathway; B, Glycerophospholipid metabolism. DGLC: D-Glucose, ALG6P: alpha-D-Glucose 6-phosphate, BTF6P: beta-D-Fructose 6-phosphate, BTFDP: beta-D-Fructose 1,6-bisphosphate, D6PGL: D-Glucono-1,5-lactone 6-phosphate, BDG6P: beta-D-Glucose 6-phosphate, D6PGC: 6-Phospho-D-gluconate, GLUA: D-Gluconic acid, R5P: D-Ribose 5-phosphate, RL5P: D-Ribulose 5-phosphate, XUL5P: D-Xylulose 5-phosphate, XUL: D-Xylulose, XYL: D-Xylose, T3P1: D-Glyceraldehyde 3-phosphate, S7P: Sedoheptulose 7-phosphate, E4P: D-Erythrose 4-phosphate, LTRY: L-Tryptophan, 12DISNGL: 1,2-Diacyl-sn-glycerol, PA: 3-sn-Phosphatidate, PPI: Diphosphate, CDPDIGLY: CDP-1,2-Diacylglycerol, PC: 1,2-Diacyl-sn-glycero-3-phosphocholine, 1ACYPHOS: 1-Acyl-sn-glycero-3-phosphocholine, GL3P: sn-Glycerol 3-phosphate, PGP: Phosphatidylglycerophosphate, SER: L-Serine, PS: Phosphatidylserine, PE: Phosphatidylethanolamine, PG: Phosphatidylglycerol, CL: Cardiolipin, GL: Glycerol

Ekwealor 发现以未脱脂大豆为氮源时比脱脂大豆更能促进 *B. megaterium* 合成赖氨酸[17]。根据所构建的代谢网络模型，在 *B. megaterium* 中存在如图 4 所示的代谢途径通路。脂肪酸被 *B. megaterium* 吸收后经过脂肪酸代谢途径生成乙酰辅酶 A，乙酰辅酶 A 反馈抑制丙酮酸向其转化，但能促进丙酮酸转化为草酰乙酸。除此之外，乙酰辅酶 A 还可通过乙醛酸循环生成大量的草酰乙酸。TCA 循环中草酰乙酸通过合成天冬氨酸后再经过赖氨酸合成途径生成赖氨酸，最终分泌到胞外。所构建模型在一定

程度上阐释了为何添加脂肪酸显著促进赖氨酸合成与分泌。

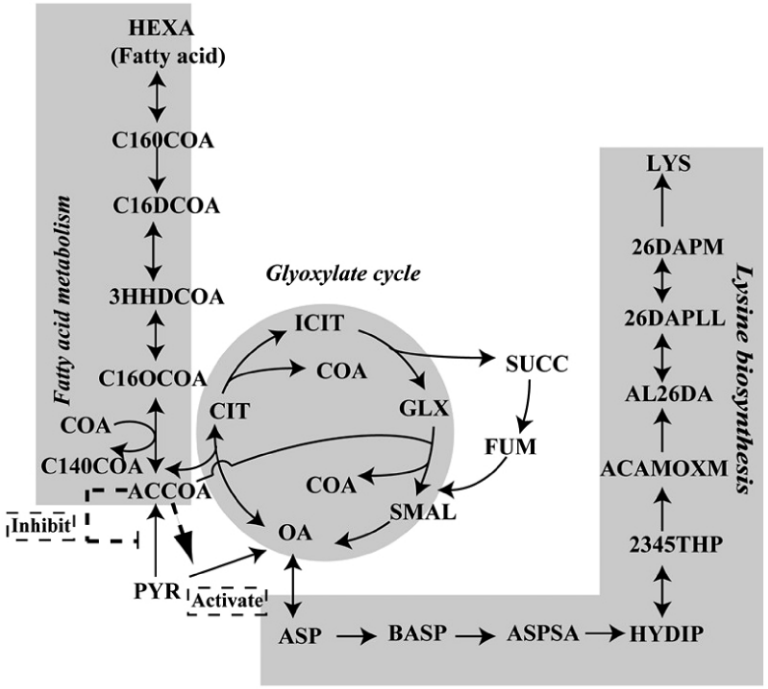

图 4　脂肪酸逐步转化为赖氨酸的通路

Fig. 4　The access of fatty acids converted into lysine in the cytosol, which involves fatty aid metabolism, glyoxylate cycle, TCA cycle and lysine biosynthesis. HEXA: Hexadecanoic acid, C160COA: Palmitoyl-CoA, C16DCOA: trans-Hexadec-2-enoyl-CoA, 3HHDCOA: (S)-3-Hydroxyhexadecanoyl-CoA, C16OCOA: 3-Oxopalmitoyl-CoA, COA: coenzyme A, C140COA: Tetradecanoyl-CoA, ACCOA: Acetyl-CoA, PYR: Pyruvate, OA: Oxaloacetate, CIT: Citrate, ICIT: Isocitrate, GLX: Glyoxylate, SMAL: (S)-malate, FUM: Fumarate, SUCC: Succinate, ASP: L-Aspartate, BASP: 4-Phospho-L-aspartate, ASPSA: L-Aspartate 4-semialdehyde, HYDIP: L-2,3-Dihydrodipicolinate, 2345THP: 2, 3, 4, 5-Tetrahydrodipicolinate, ACAMOXM: N-Acetyl-L-2-amino-6-oxopimelate, AL26DA: N6-Acetyl-LL-2, 6-diaminoheptanedioate, 26DAPLL: LL-2,6-Diaminoheptanedioate, 26DAPM: meso-2,6-Diaminoheptanedioate, LYS: L-Lysine.

3　讨论

本文通过对实验性文献进行挖掘并构建 *B. megaterium* 代谢网络模型的方法与基于基因组注释构建模型的方法的不同在于：(1)数据来源更可靠：本文模型构建所需数据来源于与 *B. megaterium* 相关的实验性文献（称之为文献组），是经过实际检验的湿数据；而后者数据是源于基因组测序的干数据；(2)数据提取方法更准确：本文采用各种文献挖掘工具和具有文献来源的数据库从文献组中获取数据而形成粗模型；而基因组注释一般通过各种专门的电子注释工具（如 KEGG Automatic Annotation Server[18]）实现粗模型数据的提取；(3)精细模型更具代表性：本文根据非物种特异的数据库（如 KEGG）以及与 *B. megaterium* 亲缘关系近的菌种（如枯草芽胞杆菌）的文献或数据库中的信息，填补粗模型的漏洞；而基于基因组信息构建模型所需的文献信息主要用于提高粗模型的特异性。因此，本文构建模型的策略是在优先保证模型可信度的基础上再提高完整性；而基于基因组注释构建模型的策略则是在保证完整性的基础上增加准确性。

采用这一策略所构建的代谢网络模型包含有丰富的中心代谢途径、维生素 B_{12} 的合成和部分氨基酸合成途径。所构建的模型在一定的条件（底物运输和基因功能）下模拟结果与实现数据比较，具有很好的吻合性。因此，可用于研究底物特异性对菌体生长的影响和预测某一基因在微生物生长和代谢网络中的生理作用，可大大降低实验操作的工作量。总之，本文的研究结果不仅提供了一种构建微生物网络模型的新策略，而且所构建的代谢网络模型在一定程度上阐述了微生物的生理功能，可以指导代

谢工程改造和工业应用。

参考文献

[1] Vary P, Biedendieck R, Fuerch T, Meinhardt F, Rohde M, Deckwer WD, Jahn D. *Bacillus megaterium*-from simple soil bacterium to industrial protein production host. *Applied Microbiology and Biotechnology*, 2007, 76 (5): 957-967.

[2] Przybulewska K, Wieczorek A, Nowak A. Isolation of microorganisms capable of styrene degradation. *Polish Journal of Environmental Studies*, 2006, 15 (5): 777-783.

[3] Crawford RL. Degradation of 3-Hydroxybenzoate by Bacteria of the Genus *Bacillus*. *Applied and Environmental Microbiology*, 1975, 30 (3): 439-444.

[4] Vary PS. Prime time for *Bacillus megaterium*. *Microbiology*, 1994, 140 (5): 1001-1013.

[5] Barg H, Malten M, Jahn M, Jahn D. Protein and Vitamin Production in *Bacillus megaterium*. Clifton: Humana Press, 2005: 205-223.

[6] Zhang J, Liu J, Shi Z, Liu L, Chen J. Manipulation of *B. megaterium* growth for efficient 2-KLG production by *K. vulgare*. *Process Biochemistry*, 2010, 45 (4): 602-606.

[7] Stammen S, Muller BK, Korneli C, Biedendieck R, Gamer M, Franco-Lara E, Jahn D. High-Yield Intra-and Extracellular Protein Production Using *Bacillus megaterium*. *Applied and Environmental Microbiology*, 2010, 76 (12): 4037-4046.

[8] Chatterjee SP, White PJ. Activities and Regulation of the Enzymes of Lysine Biosynthesis in a Lysine-excreting Strain of *Bacillus megaterium*. *Journal of General Microbiology*, 1982, 128 (5): 1073-1081.

[9] Ananiadou S, Pyysalo S, Tsujii J, Kell DB. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 2010, 28 (7): 381-390.

[10] Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 2010, 5 (1): 93-121.

[11] Kumar VS, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 2007, 8 (212).

[12] Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of metabolomics. *Briefings in Bioinformatics*, 2006, 7 (2): 140-150.

[13] Dauner M, Sauer U. Stoichiometric growth model for riboflavin-producing *Bacillus subtilis*. *Biotechnology and Bioengineering*, 2001, 76 (2): 132-43.

[14] Fürch T, Hollmann R, Wittmann C, Wang W, Deckwer WD. Comparative study on central metabolic fluxes of *Bacillus megaterium* strains in continuous culture using ^{13}C labelled substrates. *Bioprocess and Biosystems Engineering*, 2007, 30 (1): 47-59.

[15] Ulmer W, Froschle M, Jany KD. Evidence for an essential histidine residue in glucose dehydrogenase from *Bacillus megaterium* and sequence analysis of the peptides labeled with bromoacetyl pyridine. *European Journal of Biochemistry*, 1983, 136 (1): 183-94.

[16] Lach DA, Sharma VK, Vary PS. Isolation and characterization of a unique division mutant of *Bacillus megaterium*. *Journal of General Microbiology*, 1990, 136 (3): 545-553.

[17] Ekwealor IA, Obeta JAN. Studies on lysine production by *Bacillus megaterium*. *African Journal of Biotechnology*, 2005, 4 (7): 633-638.

[18] Moriya Y, Itoh M, Okuda S, Yoshizawa AC and Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 2007, 35: W182-185.

Reconstruction and analysis of *Bacillus megaterium* metabolic model based on literature study

Maoda Zhou[1,2,3], Wei Zou[1,2,3], Liming Liu[1,2,3*], Jian Chen[1,2,3]

[1] State Key Laboratory of Food Science and Technology, [2] Key Laboratory of Industrial Biotechnology, Ministry of Education, [3] Key Laboratory of Carbohydrate Chemistry and Biotechnology, Ministry of Education, Jiangnan University, Wuxi 214122, China

Abstract: **[Objective]** The aim of this study is to reconstruct the metabolic model of *Bacillus megaterium* by literature study, which would be used to elucidate physiological properties in detail and refine physiological functions. **[Methods]** We built a literature database by searching *B. megaterium* related literatures from Web of Science, PubMed, United States Patent and Trademark Office (USPTO) Patent Databases, Derwent Innovations Index and China National Knowledge Infrastructure (CNKI). Depending on this database, we extracted the functional genes, enzymes, metabolites and metabolic reactions (including transport reactions) through literature studies. Then, we filled gaps through KEGG Mapper and adding sink reactions. The operation of Matlab programs reconstructed the metabolic model (in the form of Systems Biology Markup Language) of *B. megaterium*. **[Results]** The refined metabolic network model contained 292 metabolic reactions, 378 metabolites, 220 enzymes and 217 genes. With the restrictive condition of 1.62 mmol/g cell/h of glucose uptake rate, the simulated specific growth rate was 0.089 h^{-1}, a little lower than the experimental value 0.11 h^{-1}. In addition, the accuracy of the single gene deletion simulation in pyrimidine metabolism reached 90%. **[Conclusion]** The final metabolic network model covered the biochemical information of citrate cycle, glycolysis, pentose phosphate pathway, fatty acid metabolism, vitamin B_{12} biosynthesis and amino acid biosynthesis, and reflected the effect of substrates and genes on the growth of the bacterium to a certain extent.

Keywords: *Bacillus megaterium*, literature mining, metabolic model

(本文责编:王晋芳)

Supported by the National outstanding doctorate paper author special fund (200962), by the Program for new century excellent talents in university (NCET-10-0456), by the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions and Enterprise-university-research prospective program, Jiangsu Province (BY2009112), by the Key Program of National Natural Science Foundation of China (20836003), by the Priority Academic Program Development of Jiangsu Higher Education Institutions, by the 111 Project (111-2-06) and by the Program for Advanced Talents within Six Industries of Jiangsu Province (2011-NY033)

* Corresponding author. Tel/Fax: +86-510-85197875; E-mail: mingll@jiangnan.edu.cn

Received: 12 September 2011/ Revised: 29 January 2012