酮古龙酸菌 WB0104 的全基因组分析

杨帆* 贾茜* 熊朝晖* 张笑冰* 吴洪涛* 赵颖 杨剑 朱俊萍
董杰 薛颖 孙立连 沈岩 金奇†

(病毒基因工程国家重点实验室, 北京 100176; 华北制药集团新药研究开发有限责任公司, 石家庄 050015; 国家人类基因组北方研究中心, 北京 100176. * 同等贡献. † 联系人, E-mail: zdsys@vip.sina.com)

摘要 酮古龙酸菌(*Ketogulonigenium* sp.)可以将 L-山梨糖转化为维生素 C 的前体 2-酮基-L-古龙酸, 具有重要的工业应用价值. 酮古龙酸菌 WB0104 基因组由一长度为 2765030 bp 的环形染色体和 2 个环状质粒(267968 和 242707 bp)组成, 染色体的(G+C)含量为 61.69%, 含 2727 个可读框(open reading frames, ORF), 它的复制、转录和翻译以及碳水化合物和能量代谢系统是完整的, 而修复系统不完全. WB0104 中有 640 个左右的预测 ORFs 属于转运蛋白, 占总预测基因数的四分之一左右, 远远高于目前已知的其他细菌, 推测与其适应土壤环境因素有关.

关键词 酮古龙酸菌 基因组

酮古龙酸菌(*Ketogulonigenium* sp.)是 Urbance 等人^[1]在 2001 年分类定义的一个新属, 其分类学位置为原细菌门(Proteobacteria), α -亚纲(Alpha proteobacteria), 红细菌目(Rhodobacterales), 红细菌科(Rhodobacteraceae), *Ketogulonigenium*属. 该菌首次分离自土壤, 为革兰氏阴性菌, 有鞭毛和荚膜, 兼性厌氧, 单个细胞为卵形和棒形, 直径 0.8~1.0 μm , 最佳生长温度为 27~29 $^{\circ}\text{C}$, 最佳 pH 为 7.5~8.0. 由于其具备将 L-山梨糖转化为维生素 C (vitamin c)的前体 2-酮基-L-古龙酸(2-KGA)的能力, 因此可以用于维生素 C 的生物发酵生产.

本研究以华北制药集团提供的维生素 C 生产菌株酮古龙酸菌 WB0104 为研究目标, 阐明其完整的基因组结构为研究其重要的代谢功能基因、代谢流分析以及代谢网络重建, 以及开展维生素 C 生产菌株的改造和生产工艺的改进提供最直接的依据.

1 材料方法

() 菌株. 酮古龙酸菌 WB0104 由华北制药集团提供.

() 基因组 DNA 的提取. 用常规的基因组 DNA 提取方法提取基因组 DNA, 具体方法参见《分子克隆》第 3 版.

() 全基因组序列测定及拼接. 酮古龙酸菌的全基因组序列测定采用鸟枪法(Shotgun)^[2]. 染色体 DNA 经超声打碎后, 经琼脂糖电泳, 根据以后的建库工作分别回收相应大小的 DNA 片段. 将这些片段

用 T4 DNA 聚合酶补平, 与末端去磷酸化处理的 BluescriptKS 载体相连接, 电转化大肠杆菌. 再经过质粒提取, 得到带有基因组片段的质粒模板库. 构建了 3 套插入片段大小不同的基因组文库, 插入片段的大小分别为 1000~2000, 2000~4000 以及 4000 bp 以上. 共利用了 35366 个有效序列, 完成基因组的 5.8 倍覆盖率, 然后通过随机 PCR 的方法进行物理缺口的填补以及数据拼接、验证等, 得到了全基因组核苷酸序列. 测序采用 Perkin Elmer 公司的 BigDye 荧光测序试剂盒, 数据的收集和分析采用美国应用生物系统 ABI3700 全自动序列分析仪. 酮古龙酸菌全基因组的序列组装主要使用的是由美国的华盛顿大学基因组中心的 Green 和 Ewing 等人开发的 Phred-Phrap 软件包.

在酮古龙酸菌全基因组组装过程中, 使用国际人类基因组计划所通用的质量指标 20 的标准来筛选原始序列. 相当于拼接序列中每一个碱基的精确度为 99%.

() ORF 预测与注释. 用 Glimmer 2.01 软件对 WB0104 的基因组序列进行 ORF 预测和注释. 然后将预测的 ORFs 与现有的 NR (the non-redundant protein) 和 COG (clusters of orthologous groups of proteins) 蛋白序列库作 BLASTP 比对, 从而获得功能提示, 找出与已知功能蛋白序列具有较高同源性的 ORFs. 另外, 使用 tRNAscan-SE 软件来识别基因组上的转运 RNA.

2 结果与讨论

2.1 基因组概况

全序列分析表明, 酮古龙酸菌 WB0104 染色体全

长为 2765030 bp, 环状结构; 两个环状质粒全长分别为 267968 和 242707 bp, 染色体的(G+C)含量平均为 61.69%, 含 2727 个可读框, 其中 72.4%有明确的生物学功能, 17.7%功能不明. 2727 个 ORFs 占了染色体总长的 94.4%,平均每个 ORF 大小为 957 bp, 质粒 1 的 (G+C)含量平均为 61.32%, 包含 244 个 ORF, 占质粒 1 的 91.77%; 质粒 2 的(G+C)含量平均为 62.58%, 包

含 225 个 ORF, 占质粒 2 的 94.73%(图 1, 表 1 和 2). 酮古龙酸菌的 ORFs 根据 COGs (clusters of orthologous groups of protein)可分为 17 个功能组和未知功能基因组、未分类组和 Cog 库内无比对结果组, 共 20 个组(表 2).

2.2 DNA 的复制与修复

酮古龙酸菌 WB0104 存在较完整的 DNA 复制系

图 1 酮古龙酸菌 WB0104 基因组环状图谱
最外环显示 DNA 链上顺时针转录的编码区. 第 2 环为 DNA 链上反时针转录的编码区. 每个同心环中的线表示表征位置.
第 3 和 4 环分别显示染色体(G+C)含量与 GC 偏斜((G ± C)/(G + C)). 第 5 和 6 环分别显示 rRNA 与 tRNA

表 1 酮古龙酸菌 WB0104 的基因组特征^{a)}

	染色体	质粒 1	质粒 2
大小	2765030	267968	242707
(G+C) 含量	61.69%	61.32%	62.58%
ORF 数目	2727	244	225
ORF 大小	957	1008	1022
编码百分比	94.40%	91.77%	94.73%
rRNA 数目	4	1	0
tRNA 数目	55	3	0
与已知蛋白相似的蛋白数目	2128	187	188
未知功能蛋白数目	260	21	18
保守的假定蛋白数目	23	4	0
假定蛋白数目	316	32	19

a) 未知功能蛋白: 与某一已命名的、功能尚不明确的蛋白有显著序列相似性; 保守的假定蛋白(conserved hypothetical protein): 与另一 ORF 编码的蛋白有序列相似性, 但目前无实验证据表明其被表达; 假定蛋白: 与其他已知蛋白无显著序列相似性

表 2 ORFs 按 COG 功能分类表

COG 分类号	功能分类	染色体 质粒	
J	翻译、核糖体结构及生物发生	153	5
K	转录	165	44
L	DNA 复制、重组与修复	109	19
D	细胞分裂与染色体分离	23	4
O	翻译后修饰、蛋白周转、伴侣	87	5
M	细胞外套的生物合成、外膜	126	7
N	细胞动力和分泌	68	11
P	无机离子的转运和代谢	249	90
T	信号转导系统	52	15
C	能量产生与转化	132	19
G	碳水化合物的转运与代谢	186	27
E	氨基酸的转运与代谢	343	114
F	核苷的转运与代谢	83	2
H	辅酶的代谢	97	18
I	类脂化合物的代谢	48	6
Q	次级代谢产物的生物合成、转运及分解代谢	79	7
R	预测的一般功能	315	40
S	功能未知	147	17
no COGs	未分类	261	55
COG no hits	Cog 库内无比对结果	344	58

统, 在染色体基因组中发现了启动染色体复制及 DNA 合成的 *dnaA*, *dnaB* 和 *ssb* 基因, α 型和 β 型 DNA 聚合酶(包括 α , β , $\delta\sigma$, τ , χ , $\epsilon\gamma$ 亚基), DNA 解旋酶(A, B 亚单位), ATP-依赖的 DNA 解旋酶, DNA 拓扑异构酶等. 但是, 酮古龙酸菌的 DNA 修复系统并不完善, 仅发现了 SOS 修复系统的 *recA* 和 *lexA* 基因以及 3 个甲基化错配修复基因中的 *mutL* 和 *mutS*(缺失核酸酶基因 *mutH*), 未发现切除修复系统的 *uvrABCD* 基因. 在染色体基因组还存在着与 DNA 重组相关的 *ruvA*, *B*, *C* 和 *recG* 基因.

2.3 转录和翻译

对 WB0104 基因组中与转录相关者进行分析, 其拥有与大肠杆菌相似的转录系统, 共有 165 个基因, 含有编码 RNA 聚合酶全酶($\alpha\beta\beta'\sigma$ 亚单位)的基因, 编码与转录起始位点识别相关的 RNA 聚合酶 σ 亚单位的基因, 它包括 E 因子, sigma70 与 sigma32 亚基; 编码转录延伸因子 AsnC 亚家族基因和编码转录终止子的 Rho 因子基因, 并且含有大量的具转录调节功能的转录调节蛋白(包括 MarR, ROK, LysR, GntR, AraC, tetR,

IclR, AcrR, MerR, LacI 等 10 种亚家族成员)的编码基因.

WB0104 基因组中涉及翻译、核糖体结构与合成的基因共有 153 个, 其基因组中可以找到全部 20 种基本氨基酸的氨基酰-tRNA 合成酶, 它们参与相应的氨基酸-tRNA 合成, 基因组包含了能够携带 19 种基本氨基酸的 55 个 tRNA 基因, 但在注释过程中没有发现酪氨酸-tRNA 基因. 此外, 基因组还有启动肽链合成的甲酰甲硫氨酸-tRNA-甲酰转移酶及 3 种起始因子(IF-1, IF-2, IF-3), 在肽链延长过程中发挥进位作用的延长因子 T(EF-T, 包括 EF-Ts, EF-Tu)及发挥移位作用的延长因子 G(EF-G)及延长因子 P 等, 与肽链释放相关的肽链释放因子(RF-1, RF-2, RF-3)及与多肽链翻译后加工相关的多肽去甲酰基酶、乙酰基转移酶等; 以及参与 rRNA 编码的有 3 个 rRNA 甲基化酶和一个 rRNA 甲基转移酶. 酮古龙酸菌的核糖体由 50S 与 30S 两个亚基构成, 共有 55 个相关基因, 其中包括 22 个 S 型核糖体蛋白和 27 个 L 型核糖体蛋白.

2.4 转运系统

根据分析注释的结果, 酮古龙酸菌中有 640 个左右的 ORFs(包括两个大质粒)编码转运相关蛋白, 占总预测基因数的四分之一左右. 转运蛋白基因总长度占整个基因组长度的 20% 以上, 该数值高于土壤根癌菌(*Agrobacterium tumefaciens*, 15%), 是大肠杆菌(*Escherichia coli*, 10.8%)、流感嗜血杆菌(*Haemophilus influenzae*, 9.8%)、枯草芽孢杆菌(*Bacillus subtilis*, 9.7%)、生殖器支原体(*Mycoplasma genitalium*, 10.2%)的 2 倍以上, 远远高于幽门螺旋杆菌(*Helicobacter pylori*, 5.4%)、扬氏甲烷球菌(*Methanococcus jannaschii*, 3.5%)、蓝细菌 pcc6803 (*Synechocytis pcc6803*, 4.7%)^[2]. 这些转运蛋白分属于数个超家族^[3], 如易化子超家族(major facilitator superfamily, MFS)、耐药结节分化超家族(resistance-nodulation-division superfamily, RND)、药物代谢物转运体家族(drug/metabolite transporters, DMT)、多药及毒物外排家族(multidrug and toxic compound extrusion family, MATE)和 ATP 偶联盒超家族(ATP-binding cassette, ABC)等^[1]. 按照 TC transport commission 能量偶联方式的分类原则^[2], 将酮古龙酸菌的转运蛋白分为 5 类(表 2): () 通道蛋白(channel protein); () 次级主动转运蛋白(secondary active transporter); () 初级主动转运蛋白(primary active transporter); () 基团转运蛋白(group translo

cators); () 未分类的转运蛋白(unclassified transporters).

() 通道蛋白. 酮古龙酸菌中包含 3 个通道蛋白, 都属于机械通道蛋白家族, 分别为 1 个“高电导机械通道蛋白”(large-conductance mechanosensitive channel, MscL)和 2 个“低电导机械通道蛋白”(small-conductance mechanosensitive channel, MscS)^[4,5]. MscS比MscL分子量大很多, 且MscS的压力域值只有MscL的 50%. 酮古龙酸菌中MscL大小为 160 个氨基酸, 2 个MscS大小分别为 800 和 362 个氨基酸. 细菌中机械通道蛋白的主要功能与抗高渗透压有关, 在高渗条件下可以牵拉使细胞膜发生形变并使细胞膜去极化, 从而打开或关闭某些离子通道.

() 次级主动转运蛋白. 次级主动转运蛋白是转运蛋白中数量最多, 分支最多的一类. 酮古龙酸菌中属于此类的转运蛋白数量最多的是“主要易化子超家族”MFS, 共 30 个, 其中有 26 个编码区位于染色体, 质粒上仅有 4 个. 除一个是硫酸盐特异性转运蛋白外, 都是非底物特异性蛋白(或特异性未知), 功能涉及糖类、氨基酸、核酸、药物及离子等的转运. 另外, 属于“耐药结节分化超家族”RND 的转运蛋白共有 6 个, “药物代谢物转运体家族”DMT 的共有 12 个, 这些蛋白的功能多与离子的转运及细菌耐药相关; trap 家族 8 个, 涉及四碳二羧酸类物质的转运, 以上这些基因都分布在染色体上. 酮古龙酸菌中属于此大类的转运蛋白还包括四碳二羧酸/ H^+ 同向转运蛋白、 Na^+/H^+ 逆向转运蛋白、钾离子转运蛋白、无机磷酸盐转运蛋白和砷抗性蛋白等.

() 初级主动转运蛋白. 与苜蓿根瘤菌(*Sinorhizobium meliloti*)、百脉根根瘤菌(*Mesorhizobium loti*)、土壤根癌菌(*Agrobacterium tumefaciens*)及蓝细菌 pcc6803 类似^[6], 酮古龙酸菌中有大量的ABC类的转运蛋白, 预测的ORF中属于此类蛋白的共 425 个, 占转运蛋白总数的 60%以上, 远远高于其他已经测定序列真核生物及细菌. 这些ABC类转运蛋白的底物包括糖、聚糖、氨基酸、多肽、金属离子、含金属离子的复合物、阴离子及多药(multidrug)等. 此外, 酮古龙酸菌中还含有 3 个P型ATP酶, 与阳离子转运相关.

() 基团转运蛋白. 酮古龙酸菌中含有磷酸转移酶系统(phosphotransferase system, PTS)所需的磷酸烯醇式丙酮酸(phosphoenolpyruvate, PEP)合成酶、磷酸基载体蛋白(phosphocarrier protein, HPr)和酶(enzyme)等. 另外, 在酮古龙酸菌中只发现了与果

糖相关的受体蛋白, 由此推断, 酮古龙酸菌中可能只有果糖是通过 PTS 系统来运输的.

() 未分类的转运蛋白. 包括药物外排泵及 MGTE 等.

2.5 碳水化合物代谢与转运

在酮古龙酸菌基因组中存在大量的糖转运蛋白基因提示其可利用多种碳水化合物作为能量来源, 通过生物学实验发现酮古龙酸菌可以分别利用葡萄糖、果糖、蔗糖、海藻糖等为其碳源. 基因组存在参与糖酵解的所有酶的基因, 三羧酸循环途径完全, 其磷酸戊糖途径, 是利用 L-山梨糖时的主要的提供能量的途径(图 2).

和其他细菌一样, 酮古龙酸菌的信号传导调节系统包括由感受器蛋白激酶(sensor protein kinase)和应答调节因子(response regulator)组成的双组分信号传导器(two-component signal transduction). 在酮古龙酸菌基因组中已确定了 13 个应答调节因子编码基因, 这些基因大多数毗邻于组氨酸激酶基因. 应答调节因子蛋白拥有一个相当保守的 N-端磷酸化受体区域, C-端则为 DNA 结合区. 与苜蓿根瘤菌、根癌土壤杆菌、*Caulobacter crescentus* 及假单胞铜绿菌的应答调节因子具有极高的同源性.

2.6 与 2-酮基-L-古龙酸合成相关酶的编码基因

酮古龙酸菌中包含将 L-山梨糖经 2 步催化生成 2-酮基-L-古龙酸的 L-山梨糖脱氢酶基因和 L-山梨酮脱氢酶基因. L-山梨酮脱氢酶基因与已知的 L-山梨酮脱氢酶(氧化葡萄糖酸杆菌 *Gluconobacter oxydans*)相比, 在氨基酸水平的同源性为 64%. L-山梨糖脱氢酶基因与已知的 L-山梨糖脱氢酶(氧化葡萄糖酸杆菌)比对, 在氨基酸水平的同源性为 35%. 同时酮古龙酸菌中还存在着其他与 2-酮基-L-古龙酸合成相关的酶类: L-艾杜糖脱氢酶、L-山梨糖还原酶、L-山梨酮还原酶及 2-酮基-L-古龙酸还原酶等.

酮古龙酸菌基因组序列的阐明为研究该菌的生物学背景、遗传代谢方式提供了一个新的起点, 该菌是我国用于维生素 C 生产的重要工业菌株. 以完整基因组序列为背景, 研究重要的代谢功能基因, 可以为维生素 C 生产菌株的改造和构建提供最直接的依据, 并在此基础上通过改变代谢流向, 改造代谢途径或构建新的代谢途径等手段改造菌种或优化发酵工艺, 这是工业微生物研究的一个新的方向和领域. 这

图 2 酮古龙酸菌代谢与转运概况图

灰色框表示三羧酸循环；黑色框表示磷酸戊糖途径；红色箭头表示糖酵解、脂肪酸合成、氨基酸合成。绿色表示阳离子；红色表示阴离子；黄色表示碳水化合物、醇或酸；棕色表示氨基酸或有机胺

项研究将对我国维生素 C 的生产及该领域所占据的国际地位起到决定性的作用，为国内传统工业的现代改造树立典范。

致谢 本工作为国家高技术研究发展计划(批准号: 2001AA223071)资助项目。

参 考 文 献

- 1 Urbance J W, Bratina B J, Stoddard S F, et al. Taxonomic characterization of *Ketogulonigenium vulgare* gen. nov., sp. nov. and *Ketogulonigenium robustum* sp. nov., which oxidize L-sorbose to 2-keto-L-gulononic acid. *Int J Syst Evol Microbiol*, 2001, 51: 1059—1070
- 2 Saier M H Jr, Paulsen I T. Phylogeny of multidrug transporters. *Semin Cell Dev Biol*, 2001, 12(3): 205—213 [DOI]
- 3 Paulsen I T, Sliwinski M K, Saier M H Jr. Microbial genome analyses: Global comparisons of transport capabilities based on

phylogenies, bioenergetics and substrate specificities. *J Mol Biol*, 1998, 277(3): 573—592 [DOI]

- 4 Chang G, Spencer R H, Lee A T, et al. Structure of the MscL homolog from *Mycobacterium tuberculosis*: A gated mechanosensitive ion channel. *Science*, 1998, 282(5397): 2220—2226 [DOI]
- 5 Moe P C, Blount P, Kung C. Functional and structural conservation in the mechanosensitive channel MscL implicates elements crucial for mechanosensation. *Mol Microbiol*, 1998, 28(3): 583—592 [DOI]
- 6 Wood D W, Setubal J C, Kaul R, et al. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science*, 2001, 294: 2317—2323 [DOI]
- 7 Hayashi T, Makino K, Ohnishi M, et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157: H7 and genomic comparison with a laboratory strain K-12. *DNA Res*, 2001, 8: 11—22 [DOI]

(2005-10-31 收稿, 2006-01-25 接受)

<https://engine.scichina.com/doi/10.1360/csb2006-51-8-923>