遗传 HEREDITAS (Beijing) 2012 年 4 月, 34(4): 420－430
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2012.00420

微生物必需基因的理论研究现状

叶远浓, 郭锋彪

电子科技大学生命科学与技术学院, 成都 610054

摘要: 必需基因是生物体在优化条件下生长不可缺少的基因。近年来, 对必需基因的研究已逐渐成为微生物学、基因组学和生物信息学研究领域的热点。文章首先描述了已经实验确定必需基因的微生物物种。然后, 对必需基因的理论研究现状进行了综述。从进化保守性和序列组成两方面比较必需基因和非必需基因的差异, 到必需基因的理论预测及必需基因在染色体上的分布等。最后, 对这一重要研究领域的进展进行了总结和展望。

关键词: 必需基因; 进化率; 理论预测; 染色体分布

Current status of theoretical studies on essential genes in microbes

YE Yuan-Nong, GUO Feng-Biao

School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, *China*

Abstract: Essential genes are indispensable for the survival of an organism in optimal conditions. Recently, study on essential gene is becoming a hot topic of microbiology, genomics, and bioinformatics. This paper described the experiments that determined essential genes in some microbes and the theoretical researches on essential genes were reviewed. The major content contained comparison of essential genes and non-essential genes based on information on evolutionary conservation and sequence composition, and *in silico* prediction of essential genes, and analysis of the chromosomal distributions of essential genes. Finally, related progresses were concluded and the open problems were pointed out.

Keywords: essential genes; evolution rate; theoretical prediction; chromosomal distribution

基因是遗传的基本单位, 是携带遗传信息的DNA片段, 是现代分子生物学研究的基本对象。基因分为结构基因、调节基因和操纵基因, 其中结构基因是决定合成某一种蛋白质分子结构的基因, 调节基因是调节蛋白质合成的基因, 操纵基因是控制操纵子中结构基因转录的基因。在一个生物体包含的全部基因中, 有一部分是必需基因(Essential genes), 必需基因是现代生物学研究的重中之重。必需基因是指在一定环境条件下, 维持某种生物体的生命活动所必不可少的基因。这些基因所编码蛋白质的功能被认为是生命的基础[1]。与必需基因相关的另外一个概念是最小基因集[1~3], 所谓最小基因集就是在最理想的条件下(具有足够的必需营养物质, 且没有外界环境的压力因素), 维持细胞生命活

收稿日期: 2011–04–23; 修回日期: 2011–10–29
基金项目: 国家自然科学基金项目(编号: 31071109, 60801058)资助
作者简介: 叶远浓, 在读硕士研究生, 专业方向: 生物信息学。E-mail: yyn0452@sohu.com
通讯作者: 郭锋彪, 博士, 副教授, 博士生导师, 研究方向: 微生物进化及计算基因组。E-mail: fbguo@uestc.edu.cn
网络出版时间: 2012-3-16 15:51:50
URL: http://www.cnki.net/kcms/detail/11.1913.R.20120316.1551.010.html

动能够正常运转的尽可能小的基因组合。需要注意的是离开环境单独讨论必需基因毫无意义，必须要把生物生存的外部环境作为基本的边界条件，由此不难得出，最优的生长条件对应的是最小基因集[2]。

确定必需基因的意义不仅在于它代表着生物学中一个根本的问题——了解生命起源及进化，对必需基因的理论研究有助于理解和确定生命最初共同祖先(Least common ancestor, LCA)的基因构成和生活方式[3]。同时还在于其具有巨大的应用价值，例如，因为大部分抗生素以参与基本的代谢途径的物质为靶标，致病菌的必需基因就成为潜在的抗菌药物和疫苗设计的靶标[4, 5]。

1　必需基因的实验确定

必需基因已逐渐成为基因组学和生物信息学研究领域的一个热点。随着全基因组测序的完成和基因组规模的基因失活技术的发展，确定基因组所包含的必需基因逐渐变为可能。早在 1995 年，美国科学家Itaya[6]利用诱导突变(Mutagenesis)的方法，研究了枯草芽孢杆菌(*Bacillus subtilis*)中随机选取的79个基因，结果发现其中的6个基因是不可缺失的。据此，Itaya认为枯草芽胞杆菌基因组约有 562 kb长度的DNA是必不可少的。这是对必需基因第一次较大规模的测定。1999 年，美国TIGR基因组研究所的Venter领导的研究组[7]为了调查当时测序的最小的基因组—生殖道支原体(*Mycoplasma genitalium*)中自然出现的基因集在实验室生长条件下是否已经是最小，采用全局转座子突变(Global transposon mutagenesis)的方法对生殖道支原体的基因逐个进行转座插入，并检查其存活情况，结果发现生殖道支原体可能只需要 480 个蛋白质编码基因中的265~350 个基因就可以在实验室条件下生存，其中有 100 个必需基因是功能未知的基因。这是第一次对生物体必需基因全基因组规模的确定，自此以后，越来越多的研究机构加入到了必需基因实验确定这一研究中来[8]。

在金黄色葡萄球菌(*Staphylococcus aureus* N315)的必需基因研究中，用快速鸟枪反义RNA技术(Rapid shotgun antisense RNA)结合同源比对的方法确定了658个必需基因[9, 10]，其中168个必需基因在生殖道支原体中有同源基因。后来，Ko等[11]采用等位替换突变结合比较基因组学的方法把该菌株包含的必需基因数目修改为 302 个。最近，Chaudhuri等[12]通过转座子介导的差分杂交的方法(Transposon-mediated differential hybridisation, TMDH)确定了金黄色葡萄球菌NCTC 8325 菌株(*S. aureus* NCTC 8325)的 351 个必需基因。在枯草芽孢杆菌(*B. subtilis*)基因组研究中，通过系统基因灭活(Systematic gene inactivation)技术确定了 192 个必需基因，另外又预测出了 79 个必需基因，这些必需基因分散在与细胞代谢相关的少数几个功能类中，其中大约一半和信息处理过程有关，约有 1/5 和细胞膜合成、细胞形成和分裂有关[1]。在幽门螺杆菌(*Helicobacter pylori*)中，通过转座子突变的微阵列检测技术(Microarray tracking of transposon mutants)确定了 344 个可信度很高的必需基因[13]，通过序列比较分析，发现 344 个必需基因中只有11%的必需基因在其他所有细菌中也存在同源基因，而有 55%的必需基因在一个以上的细菌中存在。在流感嗜血杆菌(*Haemophilus influenzae*)中[14]，通过高密度转座子突变(High-density transposon mutagenesis)技术确定了 478 个必需基因，其中 219 个具有已知的功能，其余 259 个则是假设的蛋白编码基因。在结核分枝杆菌(*Mycobacterium tuberculosis*)中，同样通过高密度转座子突变技术确定了 614 个必需基因[15]。通过序列比较分析，发现这些必需基因在同属的麻风分支杆菌(*Mycobacterium leprae*)中高度保守，然而，这些必需基因在其他细菌基因组中很少有同源基因。这种现象暗示了不同进化历史的物种生存所需要的必需基因具有很大区别。在大肠杆菌(*Escherichia coli*)中，通过遗传足迹法(Genetic footprinting)确定了 620 个必需基因和 3 126 个非必需基因[16]，进化分析发现，大肠杆菌的必需基因在所有细菌中都高度保守。在鼠伤寒沙门氏菌(*Salmonella typhimurium*)中，采用插入-重复突变(Insertion-duplication mutagenesis)技术确定了 257 个必需基因[17]，其中 112 个是在这次研究中新确定的，研究人员估计这些必需基因占到了整个基因组最小基因集的 52%，也就是说，沙门氏菌大约有 490 个必需基因。在肺炎链球菌(*Streptococcus pneumoniae*)中，两个研究组分别采用靶基因阻断(Targeted disruption)[18]和等位替换突变(Allelic replacement mutagenesis)[19]技术确定了 113 个和 133 个必需基因。Cameron等[20]采用

扫描近饱和转座子插入库的方式确定了霍乱弧菌(*Vibrio cholerae*C6706)临床株的 779 个必需基因。Liberati等[21]研究绿脓杆菌(*Pseudomonas aeruginosa* UCBPP-PA14)基因的必需性时，采用转座子插入突变的方法确定了 335 个必需基因。2006 年, Venter领导的研究组又采用全局转座子突变的方法对生殖道支原体进行了更为彻底的研究[8]，结果确定了 382 个必需基因，他们认为这就是最小的细菌生长所需要的最少基因数目。French等[22]则对另一种支原体(*Mycoplasma pulmonis* UAB CTIP)进行了必需基因的实验确定，通过转座子突变的方式识别出 310 个必需基因。以上 2 种支原体必需基因数目的差别应该主要是由于不同的细菌菌株的缘故，有时即使是同一个细菌种的不同菌株间也可能存在基因数目的差异。Berardinis等[23]通过单基因删除突变的方式研究了不动杆菌(*Acinetobacter baylyi*)中基因的必需性，识别了 499 个必需基因。Gallagher等[24]通过扫描转座子突变库的方式确定了弗朗西斯菌(*Francisella novicida* U112)包含的 392 个必需基因。

酿酒酵母(*Saccharomyces cerevisiae*)是目前唯一一个具有大规模必需基因实验数据的真核微生物。2002 年，Giaever等[25]在优化了酵母的生长条件下，以基因敲除的方式确定了该物种的 1 110 个必需基因。

到目前为止，对于已经完全测序的 1 000 多种细菌基因组，约有 15 种微生物(17 个菌株)的必需基因被国外研究者大规模的确定(表 1)。国内对必需基因的实验确定只局限于研究组感兴趣的少数基因。比如，上海交通大学医学院黄绍光教授的研究组以结核分枝杆菌野生临床菌株CDC1551 为母体，运用转移因子插入法建立结核分枝杆菌(*M. tuberculosis*)突变库[26]。而后应用突变株分析设计队列法、基因芯片技术与实时PCR技术，筛选在小鼠肺、豚鼠肺、活化巨噬细胞内及各种化学刺激下存活能力显著降低的突变株，这些突变株涉及的基因是结核分枝杆菌在应激状态下生存的必需基因。在这些生长条件下，他们一共发现了 47 个必需基因。

确定必需基因的实验方法有一个共同特点就是耗费巨大。另外，实验方法确定的必需基因都依赖于所采用的实验条件。这种实验条件一般都是有充足代谢供给的固体培养基来维持生长的，这与物种在野外竞争资源生长的条件不一样[27]。因此，这两种条件下的必需基因和必需基因的数目可能也不一样。实验方法还受到生物体本身的限制[28]，例如，微生物鸟枪测序联盟最近指出只有不到 1%的物种可

表 1 微生物中确定的必需基因汇总

物种名	实验方法	必需基因数目	文献
Bacillus subtilis 168	单基因灭活技术	271	[1]
Mycoplasma genitalium	转座子突变技术	382	[8]
Staphylococcus aureus N315	快速鸟枪方法反义 RNA 技术	302	[9~11]
Staphylococcus aureus NCTC 8325	转座子介导的差分杂交方法	351	[12]
Helicobacter pylori	转座子突变的微阵列检测技术	344	[13]
Haemophilus influenzae	高密度转座子突变技术	478	[14]
Mycobacterium tuberculosis	高密度转座子突变技术	614	[15]
Escherichia coli MG1655 I	遗传足迹法	620	[16]
Salmonella typhimurium	插入-重复突变	257	[17]
Streptococcus pneumoniae	靶基因阻断、等位替换变	246	[18,19]
Vibrio cholera C6706	扫描近饱和转座子插入库	779	[20]
Pseudomonas aeruginosa	转座子插入突变	335	[21]
Mycoplasma pulmonis UAB CTIP	转座子突变	310	[22]
Acinetobacter baylyi	单基因删除突变	499	[23]
Francisella novicida U112	扫描转位突变库	392	[24]
Saccharomyces cerevisiae	基因敲出	1110	[25]
Escherichia coli MG1655 II	单基因灭活技术	296	[60]
Salmonella enterica serovar Typhi	转座子突变	353	[61]

以进行实验室培养。由于实验方法的上述缺陷和局限性，致使从第一个细菌的必需基因大规模确定以来十几年的时间内，只有仅仅十几个细菌的必需基因被大规模实验确定。

2　必需和非必需基因的特征比较

为了理解必需基因的进化机制和从理论上有效地预测必需基因，对必需基因及非必需基因进行比较研究是很必要的，对其特征的比较主要包括进化保守性和组成特征 2 个方面。

由于施加在必需基因上的纯化选择更强烈，因此研究者认为必需基因比非必需基因更保守，也就是进化率更低[20]。Hirsh和Fraser最早比较了酵母基因组中必需基因和非必需基因间的进化率[29](估计的酵母蛋白的进化率，每个酿酒酵母蛋白质和线虫的同源序列之间的比较。这种比较，可以用来估计所有同时共有的序列对分歧以来的相对进化率)。他们采用了Winzeler等[30]所做的基因敲除后生长率的数据，根据生长率大小定义了一个适应度指标来定量地度量基因的必需性，为了计算进化率，则以秀丽隐杆线虫(*Caenorhabditis elegans*)作为参考物种来寻找同源基因。结果，获得了 119 个同源的必需基因和 168 个同源的非必需基因。直接比较 2 组样本的进化率，没有发现明显的差别(Welch's test, P=0.18; Wilcoxon two-sample test, P=0.18)。但是，如果按照定义的适应度各选最必需的 60 个基因和最不必需的 60 个基因，则进化率间存在显著的差别(Welch's test, P=0.00001; Wilcoxon two-sample test, P=7.2×10^{-5})。2010 年, Comas等[31]在对结核分枝杆菌人T细胞表位保守性分析的时候，也发现必需基因比非必需基因更保守。他们采用核苷酸多样性(SNPs)检测的方法对 21 株结核分枝杆菌的必需基因、非必需基因和T细胞表位的SNPs进行了统计分析，发现必需基因比非必需基因的SNPs少(Mann-Whitney U test, P<0.002)；然后又计算了 3 类基因的非冗余SNPs的非同义突变和同义突变的数值比(Nonsyn/ Syn)，发现必需基因的Nonsyn/Syn值比非必需基因的Nonsyn/Syn值要小(3 类的值分别为 1.49、1.88、1.45)，说明必需基因比非必需基因更保守。然后又计算了基因序列中非同义替换和同义替换的比值(dN/dS)，发现T细胞表位具有最好的保守性，必需基因则比非必需基因更保守(比值分别为 0.25、0.53、0.66)[31]。

Koonin的研究组比较了 3 种细菌基因组中必需基因和非必需基因间的进化率[32]。3 种细菌分别为大肠杆菌、幽门螺杆菌和脑膜炎奈瑟菌。为了获得同源序列，每个物种都采用了同种的两个菌株，分别为*E. coli* K12 和*E. coli* O157:H7; *H. pylori* 26695 和*H. pylori* J99; *Neisseria meningitidis serogroup* B strain MC58 和*N. meningitidis serogroup* A strain Z2491。其中大肠杆菌的必需基因是实验确定的，可从PEC数据库(http://www.shigen.nig.ac.jp/ecoli/pec/)获得。幽门螺杆菌和脑膜炎奈瑟菌的必需基因则是与大肠杆菌的必需基因进行序列同源比对后得到的。比较后发现, 3 个物种的必需和非必需基因间的进化率均有显著差别(Mann-Whitney U test, P<6.1×10^{-4})。为了表明以上差异不是由少数基因的偶然性偏差产生的，进行了Bootstrap抽样检验，结论依然成立。为了调查是不是特定功能类的基因间才存在偏差，把基因分成了 4 个大的功能类：信息处理和储存，细胞过程，代谢以及功能不明确的基因。发现 4 个类别间的进化率没有明显区别。对于每个类别，分别比较必需和非必需基因的进化率，结果发现 3 个类别中非同义替代率Ka都存在明显差别(Mann-Whitney U test, P<0.036)，唯一例外的是那些功能不确定类的基因。作者还分析了Hrish和Fraser[29]对酵母的必需基因和非必需基因进化率研究和这 3 种细菌间进化率研究产生了不同结果的原因(Hrish等研究酵母发现必需基因和非必需基因的进化率无显著差异，Koonin等发现在以上 3 种细菌中的必需基因和非必需基因有显著差异)，认为Hrish和Fraser的工作中非必需基因数目过少可能是一个主要原因。

对于必需基因的进化率研究还有一个是与高表达基因进化率的比较。也就是确定基因的必需性和基因表达水平，哪个因素对进化率的影响更大。Pal等[33]以白色念珠菌(*Candida albicans*)作为参照物种，在酵母基因组中比较了这两种因素和进化率的相关性。发现必需性和进化率之间有弱相关性(Pearson r=0.05, P=0.028, Spearman rank r=0.07, P=0.003)。控制表达水平后，必需性和进化率之间的相关性消失(P>0.15)。因此，他们认为是由于部分非必需基因具有较低的表达水平导致了上述微弱的相关性。Rocha和Danchin[34]在大肠杆菌(*E. coli*)和枯草芽孢杆菌(*B.*

subtilis)中比较了必需性和表达水平对进化率的影响，在该研究中，必需性没有定量的指标，而只是根据是否必需基因而采用 0 和 1 来表示必需性的大小。表达水平则采用理论指标密码子适应指数 (Codon adaptation index, CAI)。比较后发现，表达水平与进化率之间的相关性要高于必需性与进化率之间的相关性。如果采用偏相关分析控制表达水平，那么必需性与进化率之间相关系数则降低为 0.08 和 0.11。Wall等[35]对酵母基因组进行了更为系统和严格的比较。他们采用两组必需性的实验数据，一组mRNA表达数据和一组CAI度量的理论数据，两种计算进化率的理论模型，4 种相关性估计。经过比较，Wall等认为尽管必需性和进化率间的相关系数稍微低于表达水平和进化率之间的相关系数，但是两者对进化率的影响是独立起作用的。Pál等[36]和Drummond等[37]对该问题进行了总结，相对于表达水平，必需性对进化率的影响要低。但是，必需基因和非必需基因之间的进化率确实有显著的差别。

除了对进化率的研究，必需基因的进化保守性还可以表现为其他形式。Krylov等[38]在 2003 年以同源基因在 7 个真核生物中的存在作为度量基因保守或丢失的一个指标，比如 1 个基因在 6 个物种中有同源基因，则其保守率就为 6/7，丢失率就为 1/7，他们认为丢失率从某种角度上可以理解为一段时间内进化率的积分。与绝对进化率数值相比，丢失率能够有效降低统计学上的误差。与实验数据比较后发现，某同源基因丢失率和该基因被敲除后带来的致死性具有强关联，也就是说必需基因具有更广泛的系统分布。在实验确定的必需基因基础上，西北农林科技大学陶士珩研究组比较了大肠杆菌中必需基因和非必需基因的进化保守性[39]。他们用已测序的 236 个细菌基因组中有多少种细菌包含 1 个基因的同源基因作为该基因的进化保守性指标(Evolutionary retention indexes, ERI)。必需基因和非必需基因的平均ERI被发现具有显著差别(Mann–Whitney test, $P<10^{-3}$)。大约 39%的非必需基因具有极低的ERI值(<0.12)，而必需基因这样的比例只有 6.8%。大约 33%的必需基因具有极高的ERI值(>0.9)，而非必需基因这一比例为 1.65%。Bratlie等[40]最近进行了一个系统的研究，他们把ERI值高于 0.9 的基因称之为持久基因(Persistent genes)。调查后发现大多数持久基因都对应于必需基因。

窦运涛[41]曾比较了必需基因和非必需基因的单核苷酸组成。他采用相位特异性的Z曲线来代表核酸频率，每一个基因都用 9 个变量表示，对应于 9 个维空间的 1 个点，比如枯草杆菌共有 4 112 个点。为了直观的观察，应用主成分分析(Principal component analysis, PCA)方法将这些点投影到二维平面上。结果，必需基因和非必需基因都散布于整个平面。这样的结果表明必需基因和非必需基因没有出现核苷酸分布上的明显差异。对DEG数据库中其余的 8 个微生物基因组的分析也表明必需基因和非必需基因在核苷酸分布上没有明显的差异。Gong等[39]对大肠杆菌中必需和非必需基因间的氨基酸使用进行了比较，卡方分析表明两种基因间的整体密码子使用存在显著差异($P<10^{-3}$)。他们又采用PDT(Percent different test)方法逐一分析 20 种氨基酸，结果，16 种氨基酸的频率在两者间差异显著($P<0.05$)。其中，在必需基因中显著丰富的氨基酸有 5 种，在非必需基因间显著丰富的有 11 种。同时，还对 2 类基因的长度进行了比较，结果发现差异并不显著(Mann–Whitney test, $P=0.697$)。

3 必需基因的理论识别

只有随着基因组完全测序的完成和基因组规模的基因失活技术的发展，确定基因组所包含的必需基因才能成为可能。目前必需基因的确定主要依靠实验方法，而这些传统的方法大多花费巨大，甚至有些实验方法本身就存在缺陷，不能精确地确定必需基因的数目，有些实验方法的结果相互之间不能吻合。因此，国内外研究者试图利用理论方法来预测必需基因。理论预测必需基因的主要方法有比较基因组学方法、完全基于序列的方法及整合实验数据的方法。

早在 1996 年，美国NIH的科学家Koonin领导的研究组就利用比较基因组学的方法研究了细菌的最小基因集[42]。他们的研究对象是两个最先测序的细菌，生殖道支原体和流感嗜血杆菌。由于这两种细菌基因组规模都比较小，并且分别属于革兰氏阳性和革兰氏阴性菌，所以他们之间保守的基因应该就是细菌生存所必需的基因。通过详细比较，发现有 240 个生殖道支原体基因与流感嗜血杆菌基因具有

直系同源性(Ortholog)。而这种收集必需基因的过程丢失了一些为必需过程的中间步骤负责的基因，因此他们又收集了 22 个非直系同源的基因作为必需基因。这样一来，总共得到 262 个可能的必需基因。然后检查了可能的功能冗余和明显的寄生相关基因，除掉了其中的 6 个基因。剩余的 256 个就构成了细胞生存的必需的和足够的最小基因集。值得一提的是，这种方法只能识别出对大部分细菌都通用的必需基因，而不能确定某一细菌所特有的必需基因。他们得到的最小基因组的主要组成部分如下：(a)一个几乎完整的翻译系统；(b)一个基本上完整的DNA复制器；(c)一个进行DNA重组和修复的系统；(d)具有 4 个RNA聚合酶亚单位的转录装置；(e)一组参与蛋白质折叠的分子伴侣蛋白；(f)完整的无氧中间代谢途径；(g)辅酶合成途径；(h)蛋白质转运器。如果用于研究进化历史相同或相近的细菌，或者 2 个细菌基因组规模都偏大，就会造成识别的必需基因数目被高估。即使对于进化历史不同规模很小的 2 个细菌，比较基因组学方法也有它固有的缺陷。它仅仅只能识别出在进化过程保留着足够的相似度从而被确定为真正的直向同源的基因，而不能识别出物种所特有的必需基因。Gil等[3]在 2004 年也进行了最小基因集的研究，认为常用的确定必需基因 3 大实验方法都具有局限性：转座子突变技术可能会高估延缓细胞生长的非必需基因重要性而错误分为必需基因，也错失容忍转座子插入的必需基因；反义RNA抑制基因表达技术受限于抑制性 RNA在所研究生物中能得到充分表达的基因；单基因系统灭活技术的单个基因灭活检测不到冗余编码必需功能的基因。最终导致实验出来的必需基因并不是最小基因集，因为某些单独考虑时非必需的基因，在整体考虑时并不一定非必需。同时，他们通过综合方法确定了 206 个基因作为最小基因集，比Koonin[42]在 1996 年确定的基因数目少了 50 个。但是，两种最小基因集包含的大部分基因组成部分基本一致。

天津大学生物信息中心收集了通过各种方法得到的必需基因，建立了一个必需基因的数据库DEG[43]。DEG是开放数据库，全世界所有研究者都可以免费下载。DEG当前的版本是 6.8，包括了 17 种细菌和 8 种真核生物的 12 995 条必需基因记录。以DEG数据库为基础，可以通过序列比对软件Blast预测出其他细菌的必需基因。如果一个基因和DEG库中的一个或多个必需基因同源，那么这个基因对于待研究细菌的生存很可能也是必需的。近几年，国外研究者以DEG为基础识别了多个致病菌基因组中可以作为药物靶标的必需基因。比如在淋病奈瑟球菌(*Neisseria gonorrhoeae*)中，首先把已经注释好的蛋白质序列和DEG数据库中的所有必需基因的氨基酸序列逐一比对[44]，比对程序采用BlastP。而后选取满足条件(Score>100，E value<10^{-10}，and Identity>35%)的基因作为必需基因的同源序列。假设一条已知的必需基因和多个待查基因同源，那么只取相似性最高的也就是最优匹配。利用最优匹配的原理可以大大降低必需基因预测的伪正率。利用这种方法在*N. gonorrhoeae*中识别出 537 条可能的必需基因，占到了基因总数的 26%。

整合方法是近些年研究理论预测必需基因用得最多的方法，同时也是最成功的方法。所谓整合方法，主要利用各种高通量的实验数据，附加考虑进化保守特征和序列特征。Roberts等[45]曾综述了利用整合方法预测必需基因的早期工作，这些方法使用的实验数据主要有功能分类数据、基因表达数据和蛋白质相互作用数据。组合使用各种特征并采用机器学习方法进行分类是目前预测必需基因的趋势。Plaimas等[46]采用了 33 种特征，使用支持向量机来训练模型，跨物种的交叉验证表明预测的AUC(Area under the receiver-operator-curve)分数值为 75%~81%；Deng等[47]考虑了 3 类特征：(1)序列固有的特征，如GC含量和蛋白质长度；(2)从基因组序列衍生的特征，如亚细胞位置、密码子倾向性、进化保守性等，已有成熟的相应计算软件计算这些特征；(3)高通量实验数据，如基因表达芯片数据等。他们采用朴素贝叶斯分类、逻辑回归、C4.5 决策树、CN2 规则等 4 类分类器。利用朴素贝叶斯方法筛选过滤后得到如表 2 所示的 13 种特征。分别在 2 个细菌*E. coli*和*A. baylyi* ADP1 基因组内进行了种内的交叉检验，这 2 种细菌的必需基因都已经在全基因组内通过实验确定，种内的十重交叉验证表明AUC的变化范围为 0.86~0.93，跨物种的交叉验证的AUC分数为 0.69~0.89。与以往的同类研究相比，Deng等的预测精度更高，而且他们表明了可以把基于一个物种训练的模

表 2 用于必需基因预测的 13 种重要特征

序列固有特征 (排序)	序列衍生特征(排序)	实验特征(排序)
密码子偏好性(2)	富集域得分 (1)	基因表达差异(12)
疏水性(3)	进化得分 (4)	共表达网络瓶颈节点(7)
氨基酸长度(5)	亚细胞定位：细胞质 (6)	共表达网络交叉节点(9)
芳香性 (8)	亚细胞定位：胞外 (10)	
	横向同源 (11)	
	亚细胞定位：内膜(13)	

注：括号内的数字表示分类准确率的相对贡献，其中数字 1 表示最重要的特征，数字 13 表示最不重要的变量。

型成功转移到另一个远缘物种。

单纯基于基因组或蛋白质组序列提取特征来预测必需基因是理论预测的最理想化的途径，因为这种方法不像比较基因组学方法那样需要近缘物种的相关实验数据和很大的计算量，也不像整合方法那样需要物种本身的各种组学实验数据。第一个完全基于序列的必需基因识别算法由美国耶鲁大学的Gerstein领导的研究人员开发，专用于酵母基因组[28]。他们选择了 14 个和必需基因相关的特征作为变量，包括亚细胞位置信号、密码子适应性、有效密码子数、GC含量，基因长度、蛋白质的整体疏水性、二级结构单元螺旋的含量、稀有氨基酸百分比、与终止密码子相似的密码子所占的百分比等。首先采用Bayes网络测量每个变量和必需性之间的相关性。然后用机器学习分类器来学习参数，采用酿酒酵母(*S. cerevisiae*)已知的必需基因来训练，随后把训练好的模型应用于其近缘物种*Saccharomyces mikatae*。最终，他们采用同源作图和基因敲除的方法来检验结果的准确性。该算法针对真核生物酿酒酵母设计，很难推广到细菌和高等真核生物基因组研究中去。算法的准确率的绝对大小依赖于所选用的概率阈值，比如，在概率阈值约等于 0.5 时，以 4 648 个基因作训练集采用十重交叉验证可以正确识别出 88 个必需基因，丢掉 875 个。同时正确识别出 3 646 个非必需基因，把剩余的 39 个非必需基因误判为必需基因。

Baran和Ko[48]试图只利用序列组成来预测大肠杆菌的必需基因，他们首先提出了一个基于窗口的双核苷酸偏差度量指标，认为与物种平均组成有偏差的基因只能由 2 种情况产生：最新转移的基因和非常古老的基因。因此在检查出大肠杆菌所有具有组成偏性的基因后，把它们逐一和肠道沙门氏菌(*Salmonella enterica*)进行比较。如果没有发现具有显著相似性的序列(E value>0.1)或者具有可移动遗传元件，则认为是正确预测的水平转移基因。如果比对的E值小于阈值，则认为是错误预测的水平转移基因。而错误预测的水平转移基因主要对应于古老的蛋白，因此他们很可能就是必需基因。20 个具有最高组成偏差(用卡方分布测试度量)片段对应于 19 个基因，其中 4 个具有可移动遗传元件而被认为是真正的水平转移基因。14 个具有明显的相似性构成了错误预测的水平转移基因，其中有 8 个对应于实验验证的必需基因。作为对比，随机发现 1 个必需基因的概率只有 6.5%(等于必需基因数目除于基因总数)。因此，高偏差的组成可以认为是必需基因的一个强烈标志。因此，这项工作给那些试图只从序列组成预测必需基因的研究者带来了希望。

吉林大学胡成全教授的研究组单纯采用序列特征对大肠杆菌和枯草杆菌中的必需基因进行了预测[49]。从 35 个原始变量出发，采用逐步回归法筛选出 7 个变量，分别为密码子适应性指数CAI、基因长度、单氨基酸Arg和Val的含量、链偏差、疏水性值、第三位碱基A的含量。利用三层前馈神经网络对大肠杆菌的必需基因进行分类，采用贝叶斯归一化学习算法时总体(必需基因和非必需基因)分类准确率达到 66.6%，必需基因的分类准确率为 72.%。当采用学习率可变动的前馈动量算法时，总体分类准确率为 49.2%，必需基因分类准确率为 60.0%。枯草芽孢杆菌的分类准确率要高很多，无论采用哪种学习算法分类准确率都高于 60%。

4 必需基因在染色体上的分布

DNA复制是一个不对称的过程，不对称的一个体现是复制链分为前导链和滞后链。很早以前，研究者就发现了基因数目在两条复制链的不平等分

布[50,51]。与滞后链相比，前导链上分布的基因数目更多，人们过去认为这种不对称应该归因于高表达基因，因为以核糖体蛋白编码基因为代表的高表达基因在两条复制链上的密度差别更显著[52]。后来，Rocha和Danchin[53]等指出必需基因才是链间基因方向偏差的驱动力，他们先是在大肠杆菌和枯草芽孢杆菌这两个具有实验验证基因的物种间进行分析，比较了必需且高表达基因、必需且非高表达基因、非必需且高表达基因以及非必需非高表达基因 4 组样本在两个复制链上的分布情况，结果发现，必需基因的分布比高表达基因的分布偏差更显著(P<0.01)。在同一年，他们又把该研究扩展到低GC含量的厚壁菌和 γ 变形菌[52]。由于这些细菌没有实验确定的必需基因，就通过同源比对的方法来预测必需基因，比较后发现，必需基因驱动基因方向偏差的结论在这些物种中同样成立。最近，Lin等[54]进一步指出，是某些特定功能类的必需基因，而不是全部必需基因导致了复制链间的基因密度偏差。他们研究了 10 种具有实验必需基因数据的细菌，与Rocha的结论相同，Lin等在 10 种细菌中观察到必需基因优先地分布于前导链上。把必需基因和非必需基因按照COG注释分为不同的功能类后，Lin等[54]进一步发现，只有某些确定功能类的必需基因才有链偏差。这些功能类依次为：信息储存和处理相关的功能类J(COG缩写符号)、K、L，以及功能类D(细胞循环控制)、M(细胞壁生物合成)、O(后翻译修饰)、C(能量产生和转换)、G(糖类转运和代谢)、E(氨基酸转运和代谢)、以及F(核苷酸转运和代谢) 等共 10 个功能类，其他功能类的必需基因大多数在 2 个复制链间不存在显著的方向偏差(Paired t-test, P>0.01)。

除了 2 条复制链间的偏差，必需基因在同一条复制链的不同位置间也存在偏差。具体地说，必需基因和高表达基因在复制起始附近的区域要比在终止附近的区域分布密度高[55]。在大肠杆菌中，这种分布趋势尤其明显，可以近似地呈线性分布。随着离复制起始点的距离增加，必需基因的密度呈以线性逐渐降低[55]，这种分布模式被Couturier等[56]称之为复制过程的剂量效应，该效应是为了满足必需基因及其他与转录、翻译过程相关基因对表达量的高要求。最近，Guo等[57]研究了多染色体致病菌 *Burkholderia cenocepacia*中的必需基因分布情况，在用比较基因组学方法预测出全部必需基因后，发现长度最短的 3 号染色体包含的必需基因比长一倍的 2 号染色体上的必需基因还要多。分析后发现，产生这种现象的原因是该物种染色体间存在着转移。通过染色体转移, 1 号染色体上的一部分必需基因转移到了 3 号染色体上并在之后的进化中稳定了下来。

有关必需基因的最新实验报道还包括：(1)2010年底，Chen等[58]发现在果蝇基因组中，近期产生的新基因中和古老的基因中包含的必需基因比例大致相等，这样的结果表明有很大一部分新基因可以快速的进化出必需的功能并参与物种的发育；(2)2011年初, Nichols等[59]提出“条件必需基因”的概念，指在特定的生长条件下所必需的基因，这些基因在染色体的不同位置同样有着偏差的分布。

5　结语与展望

必需基因的研究正逐渐成为微生物学、遗传学、合成生物学等相关学科研究的热点。除了通过实验确定微生物生长必需的基因，相关的理论研究也很广泛。主要的理论研究为(1)必需基因和非必需基因的对比；(2)必需基因的理论预测；(3)必需基因在染色体上的分布。

作为基因组学和生物信息学研究领域的一个热点，目前对必需基因的研究已逐渐取得一定的成果，但仍有一些基本问题需要继续探索，比如：(1)必需性究竟对进化率的变化起多大的作用，这一影响在不同物种间是否一致，真核生物和原核生物在这方面是否有一致的规律；(2)在少数物种中已被证明高度有效的整合预测算法对于更广泛的物种是否依然有效；(3)单纯从序列提取特征而不采用实验数据是否可以高度准确的预测必需基因，如果可以，那么发展彻底不依赖于实验数据的自训练算法将是一个极富挑战性也是很有前途的工作；(4)必需基因在染色体的复制链间及链内的分布偏差是否是一个普遍的现象，也就是说，是不是绝大多数物种的前导链具有比滞后链更多的必需基因数目，复制起点区域具有比终止区域具有更多的数目；(5)真核生物是否具有链间分布偏差和链内分布偏差的特性。除了以上的问题，随着研究的深入还会出现越来越多有趣的问题，这些问题的完全解决需要相关研究人员，尤其是理论研究者和实验科学家的共同努力。

参考文献(References):

[1] Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Débarbouillé M, Dervyn E, Deuerling E, Devine K, Devine SK, Dreesen O, Errington J, Fillinger S, Foster SJ, Fujita Y, Galizzi A, Gardan R, Eschevins C, Fukushima T, Haga K, Harwood CR, Hecker M, Hosoya D, Hullo MF, Kakeshita H, Karamata D, Kasahara Y, Kawamura F, Koga K, Koski P, Kuwana R, Imamura D, Ishimaru M, Ishikawa S, Ishio I, Le Coq D, Masson A, Mauël C, Meima R, Mellado RP, Moir A, Moriya S, Nagakawa E, Nanamiya H, Nakai S, Nygaard P, Ogura M, Ohanan T, O'Reilly M, O'Rourke M, Pragai Z, Pooley HM, Rapoport G, Rawlins JP, Rivas LA, Rivolta C, Sadaie A, Sadaie Y, Sarvas M, Sato T, Saxild HH, Scanlan E, Schumann W, Seegers JFML, Sekiguchi J, Sekowska A, Séror SJ, Simon M, Stragier P, Studer R, Takamatsu H, Tanaka T, Takeuchi M, Thomaides HB, Vagner V, van Dijl JM, Watabe K, Wipat A, Yamamoto H, Yamamoto M, Yamamoto Y, Yamane K, Yata K ,Yoshida K, Yoshikawa H, Zuber U, Ogasawara N. Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci USA*, 2003, 100(8): 4678–4683.DOI

[2] Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol*, 2003, 1(2): 127–136. DOI

[3] Gil R, Silva FJ, Peretó J, Moya A. Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev*, 2004, 68(3): 518–537. DOI

[4] Hu WQ, Sillaots S, Lemieux S, Davison J, Kauffman S, Breton A, Linteau A, Xin CL, Bowman J, Becker J, Jiang B, Roemer T. Essential gene identification and drug target prioritization in *Aspergillus fumigatus*. *PLoS Pathog*, 2007, 3(3): e24. DOI

[5] Haselbeck R, Wall D, Jiang B, Ketela T, Zyskind J, Bussey H, Foulkes JG, Roemer T. Comprehensive essential gene Identification as a platform for novel antiinfective drug discovery. *Curr Pharm Des*, 2002, 8(13): 1155–1172. DOI

[6] Itaya M. An estimation of minimal genome size required for life. *FEBS Lett*, 1995, 362(3): 257–260. DOI

[7] Hutchison CA III, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter1 JC. Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, 1999, 286(5447): 2165–2169. DOI

[8] Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA III, Smith HO, Venter JC. Essential genes of a minimal bacterium. *Proc Natl Acad Sci USA*, 2006, 103(2): 425–430.DOI

[9] Ji YD, Zhang B, van Horn SF, Warren P, Woodnutt G, Burnham MKR, Rosenberg M. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science*, 2001, 293(5538): 2266–2269.DOI

[10] Forsyth RA, Haselbeck RJ, Ohlsen KL, Yamamoto RT, Xu H, Trawick JD, Wall D, Wang LS, Brown-Driver V, Froelich JM, C KG, King P, McCarthy M, Malone C, Misiner B, Robbins D, Tan ZH, Zhu ZY, Carr G, Mosca DA, Zamudio C, Foulkes JG, Zyskind JW. A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol Microbiol*, 2002, 43(6): 1387–1400. DOI

[11] Ko KS, Lee JY, Song JH, Baek JY, Oh WS, Chun JK, Yoon HS. Screening of Essential genes in *Staphylococcus aureus* N315 using comparative genomics and allelic replacement mutagenesis. *J Microbiol Biotechnol*, 2006, 16(4): 623–632. DOI

[12] Chaudhuri RR, Allen AG, Owen PJ, Shalom G, Stone K, Harrison M, Burgis TA, Lockyer M, Jorge GL, Foster SJ, Pleasance SJ, Peters SE, Maskell DJ, Charles IG. Comprehensive identification of essential *Staphylococcus aureus* genes using transposon-mediated differential hybridisation (TMDH). *BMC Genomics*, 2009, 10(1): 291. DOI

[13] Salama NR, Shepherd B, Falkow S. Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J Bacteriol*, 2004, 186(23): 7926–7935. DOI

[14] Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N, Mekalanos JJ. A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc Natl Acad Sci USA*, 2002, 99(2): 966–971. DOI

[15] Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol*, 2003, 48(1): 77–84. DOI

[16] Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabási AL, Oltvai ZN, Osterman AL. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol*, 2003, 185(19): 5673–5684. DOI

[17] Knuth K, Niesalla H, Hueck CJ, Fuchs TM. Large-scale identification of essential *Salmonella* genes by trapping lethal insertions. *Mol Microbiol*, 2004, 51(6): 1729–1744. DOI

[18] Thanassi JA, Hartman-Neumann SL, Dougherty TJ, Dougherty BA, Pucci MJ. Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. *Nucleic Acids Res*, 2002, 30(14): 3152–3162. DOI

[19] Song JH, Ko KS, Lee JY, Baek JY, Oh WS, Yoon HS, Jeong JY, Chun J. Identification of essential genes in *Streptococcus pneumoniae* by allelic replacement mutagenesis. *Mol Cells*, 2005, 19(3): 365–374. DOI

[20] Cameron DE, Urbach MJ, Mekalanos JJ. A defined transposon mutant library and its use in identifying motility genes in *Vibrio cholerae*. *Proc Natl Acad Sci USA*, 2008, 105(25): 8736–8741. DOI

[21] Liberati NL, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei T, Ausubel FM. An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc Natl Acad Sci USA*, 2006, 103(8): 2833–2838. DOI

[22] French CT, Lao P, Loraine AE, Matthews BT, Yu HL, Dybvig K. Large-scale transposon mutagenesis of *Mycoplasma pulmonis*. *Mol Microbiol*, 2008, 69(1): 67–76. DOI

[23] de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, Samair S, Lechaplais C, Gyapay G, Richez C, Durot M, Kreimeyer A, Fèvre FL, Schächter V, Pezo V, Döring V, Scarpelli C, Médigue C, Cohen GN, Marlière P, Salanoubat M, Weissenbach J. A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Mol Syst Biol*, 2008, 4: 174. DOI

[24] Gallagher LA, Ramage E, Jacobs MA, Kaul R, Brittnacher M, Manoil C. A comprehensive transposon mutant library of *Francisella novicida*, a bioweapon surrogate. *Proc Natl Acad Sci USA*, 2007, 104(3): 1009–1014. DOI

[25] Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau–Danila A, Anderson K, André B, Arkin AP, Astromoff A, El Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo CY, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Petra RMa, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sharon SM, Storms RK, Strathern JN, Valle G, Voet M, VolckaertG, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang YH, Yen G, Youngman E, Yu KX, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 2002, 418(6896): 387–391. DOI

[26] 程齐俭. 结核分支杆菌宿主体内及体外生长必需基因的研究[博士学位论文]. 上海: 上海交通大学, 2007. DOI

[27] Fang G, Rocha E, Danchin A. How essential are nonessential genes? *Mol Biol Evol*, 2005, 22(11): 2147–2156. DOI

[28] Seringhaus M, Paccanaro A, Borneman A, Snyder M, Gerstein M. Predicting essential genes in fungal genomes. *Genome Res*, 2006, 16(1): 1126–1135. DOI

[29] Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. *Nature*, 2001, 411(6841): 1046–1049. DOI

[30] Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Véronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu KX, Zimmermann K, Philippsen P, Johnston M, Davis RW. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, 1999, 285(5429): 901–906. DOI

[31] Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet*, 2010, 42(6): 498–503. DOI

[32] Jordan IK, Rogozin IB, Wolf YI, Koonin EK. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*, 2002, 12(6): 962–968. DOI

[33] Pál C, Papp B, Hurst LD. Genomic function: Rate of evolution and gene dispensability. *Nature*, 2003, 421(6922): 497–498. DOI

[34] Rocha EP, Danchin A. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol*, 2004, 21(1): 108–116. DOI

[35] Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA*, 2005, 102(15): 5483–5488. DOI

[36] Pál C, Papp B, Lercher MJ. An integrated view of protein evolution. *Nat Rev Genet*, 2006, 7(5): 337–348. DOI

[37] Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence

evolution. *Cell*, 2008, 134(2): 341–352. DOI

[38] Krylov DM, Wolf YI, Rogozin IB, Koonin EV. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res*, 2003, 13(10): 2229–2235. DOI

[39] Gong XD, Fan SH, Bilderbeck A, Li MK, Pang HX, Tao SH. Comparative analysis of essential genes and nonessential genes in *Escherichia coli K12*. *Mol Genet Genomics*, 2008, 279(1): 87–94. DOI

[40] Bratlie MS, Johansen J, Drabløs F. Relationship between operon preference and functional properties of persistent genes in bacterial genomes. *BMC Genomics*, 2010, 11(1): 71. DOI

[41] 窦运涛. 基于必需基因数据库的微生物必需基因的分析. 天津理工大学学报, 2006, 22(2): 9–13. DOI

[42] Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA*, 1996, 93(19): 10268–10273. DOI

[43] Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes. *Nucleic Acids Res*, 2004, 32(Database issue): D271–D272. DOI

[44] Barh D, Kumar A. In silico identification of candidate drug and vaccine targets from various pathways in *Neisseria gonorrhoeae*. *In Silico Biol*, 2009, 9(4): 225–231. DOI

[45] Roberts SB, Mazurie AJ, Buck GA. Integrating genome-scale data for gene essentiality prediction. *Chem Biodivers*, 2007, 4(11): 2618–2630. DOI

[46] Plaimas K, Eils R, König R. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst Biol*, 2010, 4(1): 56. DOI

[47] Deng JY, Deng L, Su SC, Zhang ML, Lin XD, Wei L, Minai AA, Hassett DJ, Lu LJ. Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res*, 2011, 39(3): 795–807. DOI

[48] Baran RH, Ko H. Detecting horizontally transferred and essential genes based on dinucleotide relative abundance. *DNA Res*, 2008, 15(5): 267–276. DOI

[49] 李赫. 基于人工神经网络的必需基因预测研究[硕士学位论文]. 吉林: 吉林大学, 2008. DOI

[50] McLean MJ, Wolfe KH, Devine KM. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol*, 1998, 47(6): 691–696. DOI

[51] Mrázek J, Karlin S. Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci USA*, 1998, 95(7): 3720–3725. DOI

[52] Rocha EPC, Danchin A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res*, 2003, 31(22): 6570–6577. DOI

[53] Rocha EP, Danchin A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet*, 2003, 34(4): 377–378. DOI

[54] Lin Y, Gao F, Zhang CT. Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochem Biophys Res Commun*, 2010, 396(2): 472–476. DOI

[55] Rocha EP. The replication-related organization of bacterial genomes. *Microbiology*, 2004, 150(Pt6): 1609–1627. DOI

[56] Couturier E, Rocha EPC. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol Microbiol*, 2006, 59(5): 1506–1518. DOI

[57] Guo FB, Ning LW, Huang J, Lin H, Zhang HX. Chromosome translocation and its consequence in the genome of *Burkholderia cenocepacia* AU-1054. *Biochem Biophys Res Commun*, 2010, 403(3–4): 375–379. DOI

[58] Chen SD, Zhang YE, Long MY. New genes in Drosophila quickly become essential. *Science*, 2010, 330(6011): 1682–1685. DOI

[59] Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA. Phenotypic landscape of a bacterial cell. *Cell*, 2011, 144(1): 143–156. DOI

[60] Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, 2006, 2: 2006.0008. DOI

[61] Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK. Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. *Genome Res*, 2009, 19(12): 2308–2316. DOI