Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.160407
March 25, 2017, 33(3): 324−330

·科学意义·

新基因起源：从自然进化到人工设计

王千[1,2]，程健[2]，江会锋[2]

中国科学院天津工业生物技术研究所，天津 300308

王千, 程健, 江会锋. 新基因起源：从自然进化到人工设计. 生物工程学报, 2017, 33(3): 324–330.
Wang Q, Cheng J, Jiang HF. Origin of new genes: from evolution to design. Chin J Biotech, 2017, 33(3): 324–330.

摘 要：生命体系历经40多亿年的自然进化，创造了无数丰富多彩的功能基因，保障了生命体系的传承与繁荣。然而生命体系的自然进化历程极其缓慢，新的功能基因产生需要数百万年时间，无法满足快速发展的工业生产需求。利用合成生物学技术，研究人员可以依据已知的酶催化机理和蛋白质结构进行全新的基因设计与合成，按照工业生产需求快速创造全新的蛋白质催化剂，实现各种自然界生物无法催化的生物化学反应。尽管新基因设计技术展现了激动人心的应用前景，但是目前该技术还存在设计成功率不高、酶催化活性较低、合成成本较高等科技挑战。未来随着合成生物学技术的快速发展，设计、改造、合成和筛选等技术将融合为一体，为新基因设计与创建带来全新的发展机遇。

关键词：基因合成，理性改造，基因设计，新酶设计

Origin of new genes: from evolution to design

Qian Wang[1, 2], Jian Cheng[2], and Huifeng Jiang[2]

Tianjin Institute of Industrial Biotechnology, *Chinese Academy of Sciences*, *Tianjin* 300308, *China*

Abstract: Life system has created rich and colorful genes, to protect the inheritance and prosperity after more than 4 billion years of natural evolution. However, the natural evolution is an extremely slow process, and the origin and evolution of new gene with new function often takes millions of years. Therefore, natural evolution alone cannot meet the rapid development of industrial biotechnological production needs. Using synthetic biology techniques, researchers can design and synthesize new genes based on the known enzyme catalysis mechanism and protein structure according to industrial production requirements, and create various biochemical reactions that cannot be catalyzed by natural living organisms.

Received: October 29, 2016; **Accepted:** December 16, 2016
Supported by: National Natural Science Foundation of China (No. 31300077).
Corresponding author: Huifeng Jiang. Tel: +86-22-24828732; E-mail: jiang_hf@tib.cas.cn
国家自然科学基金 (No. 31300077) 资助。
网络出版时间：2016-12-27 网络出版地址：http://www.cnki.net/kcms/detail/11.1998.Q.20161227.1445.003.html

Although the new gene design technology shows exciting application prospects, there are now still many scientific and technological challenges, such as low success rate of design, low catalytic activity and high synthesis cost. With the rapid development of synthetic biology, the design, transformation, synthesis, screening and other technologies will be integrated into a mature technological process for the new gene design.

Keywords: gene synthesis, rational engineering, gene design, new enzyme design

基因是生命体的基本功能单元，由 DNA 编码，可转录成 mRNA，并翻译成蛋白质行使功能。各种精彩纷呈的基因功能，如光合作用、新陈代谢、细胞分裂、个体发育等无不彰显了基因的神奇和生命的无穷魅力。在生命亿万年的进化历程中，新功能新基因的呈现是生命体环境适应性进化的基本保障。正是由于生物进化过程中不断产生的全新功能基因，使得生物具有应对变化莫测的地球环境的本领，从而在地球上生存了数十亿年，并逐步改造地球环境，直至今天形成了我们看到的适合人类居住的家园。

1 新基因的自然起源进化

在生命从简单到复杂的进化历程中，基因数量展示了由少到多的变化趋势，如简单的原核生物一般基因数量在几百到几千，而复杂高等的真核生物则有多达几万个的功能基因。然而新基因从何而来，新功能如何产生，又是如何参与生物进化过程等科学问题一直是困扰进化生物学家的难题。早在 1970 年日本进化学家 Ohno 首次系统阐述了新基因如何通过基因重复起源，并且认为基因重复是新基因产生的主要分子机制[1]。1993 年华人进化学家龙漫远教授首次用实验方法发现并解析了第一个由两个不同基因片段嵌合的新基因[2]，证明了基因嵌合起源的分子机理。2000 年之后，生物学研究进入了基因组时代，迅猛发展的基因组技术和庞大的基因组数据为基因起源与进化研究提供了绝佳的机遇。通过比较分析近缘物种的基因组序列，研究人员发现了多种新基因起源的分子机制，包括基因重复、基因分裂与融合、基因转座、基因横向迁移、基因嵌合和基因从头起源(从非编码区起源) 等[3]。新基因起源与进化的基础理论得到了前所未有的快速发展。

基因重复后其中一个拷贝积累突变并产生新功能的起源机制已经广为人知，或者不同功能来源的基因片段组合然后产生全新的功能基因的机制也被研究得相当清楚，但是新基因如何由非编码区起源，新功能如何从无到有的创造一直被认为是小概率事件，其起源进化机制也知之甚少。王文等以模式生物果蝇为研究对象，系统分析了黑腹果蝇基因组中的新基因起源机制[4]。他们发现除了基因重复起源机制之外，还有 12%的新基因是从头起源的，表明从头起源新基因在物种进化过程中占据了很高的比例，可能发挥了很重要的生物功能。其后，李丹等以酿酒酵母为模型深入研究了从头起源新基因的功能，发现从头起源新基因 MDF1 在酵母有性生殖和营养生长过程中发挥了重要作用，并且提高了该物种在多变环境下的适应能力[5-6]。尽管从头起源新基因是从没有生物功能的非编码区域产生的，但是其在物种适应性进化过程中具有不可替代的作用[7]。

由于所有蛋白质都需要通过核糖体与其信使 RNA (mRNA) 结合进行翻译，因此利用高通量测序技术检测被核糖体结合的 mRNA 的原理，研究人员理论上可以观察到基因组中所有编码蛋白质的基因。利用该技术，Carvunis 等发现除广为人知的 6 000 多个编码基因之外，酿酒酵母的基因组中还有 1 900 多个新的编码基因[8]。Stern-Ginossar 等采用同样的技术，解析了一个人源病毒基因组中所有的蛋白质翻译事件，发现了上百个未被注释的新编码基因[9]。同样地，刘晓秋等在经典的模式病毒 Lambda 噬菌体中，发现了 50 多个新编码基因，占到了过去几十年该病毒中已知基因总数的 80%[10]。在这些新发现的编码基因当中，部分基因已经证实有翻译的蛋白质，部分基因在近缘物种间非常保守，很可能是有生物学功能的。而且有意思的是，这些新的编码基因绝大多数都是新近起源或从头起源的新基因[8]。因此，这种新的基因发掘技术完全颠覆了传统编码基因的研究策略，为新基因的进化研究开辟了一片全新的天地。

2 新基因人工设计研究进展

新基因自然进化起源都是以百万年为单位[11]，新基因产生速率极其缓慢，远远无法满足日益增长的工业生产的需求。基于数学、物理、计算科学、工程科学与生命科学的深度融合，合成生物学推动了从认识生命到设计生命的质的变革，带来了生命科学领域的第三次革命[12-13]。合成生物学为新基因研究带来颠覆性的理念和方法。在 DNA 合成技术的武装下，人工设计与合成全新的功能基因成为了可能。依据有机化学反应原理和已有的蛋白质结构模板从头设计新的酶催化剂已经获得成功[14-15]。未来随着新基因设计技术进步，人类可以根据工业生产的需求创造出完全不同于自然生命体系的具有新基因新功能的“人造生命”[16-18]，这将为生命科学研究带来前所未有的变革。

新基因设计是指按照研究者的意愿，设计和制造出自然界不存在的、具有特定生物学功能的全新蛋白质编码基因。1988 年 Regan 等首次人工设计了可成功折叠的蛋白质[19]，但只有少数成功折叠的蛋白质具有生物活性[20]。近期刘海燕等通过能量优化和精巧的高通量筛选设计，进一步提高了设计可折叠蛋白质的成功率[21]。然而蛋白质的成功折叠并不意味着具有生物活性。为此，Baker 等开发了一套基于 Rosetta 算法的新酶设计流程，设计了大量具有生物功能的新酶[22-23]。研究人员首先构建酶催化过程中氨基酸残基与过渡态底物相互作用的量化模型[24]，搜索与模型匹配的已知蛋白质结构框架[25-27]，并将模型与蛋白质结构框架进行整合，优化底物亲和力、结合电势能和结构稳定性等[28]，最终合成表达并通过实验筛选出具有生物功能的蛋白质 (图 1A)。很多具有重要生物功能的新酶被设计出来 (表 1)，如催化羟醛缩合反应 (Retro-aldol reaction)[22]、Kemp 消除反应 (Kemp elimination reaction，图 1B)[14]、狄尔斯-阿尔德反应 (Diels-Alder reaction)[23]。同时生物代谢途径也可以借助于新酶设计进行重新构建，例如 Siegel 等利用新酶设计创建了以二氧化碳为原料合成羟基丙酮的关键催化酶，并在大肠杆菌中构建了该合成途径[29]。这些蛋白酶的成功设计充分证明新酶设计策略具有巨大发展潜力。

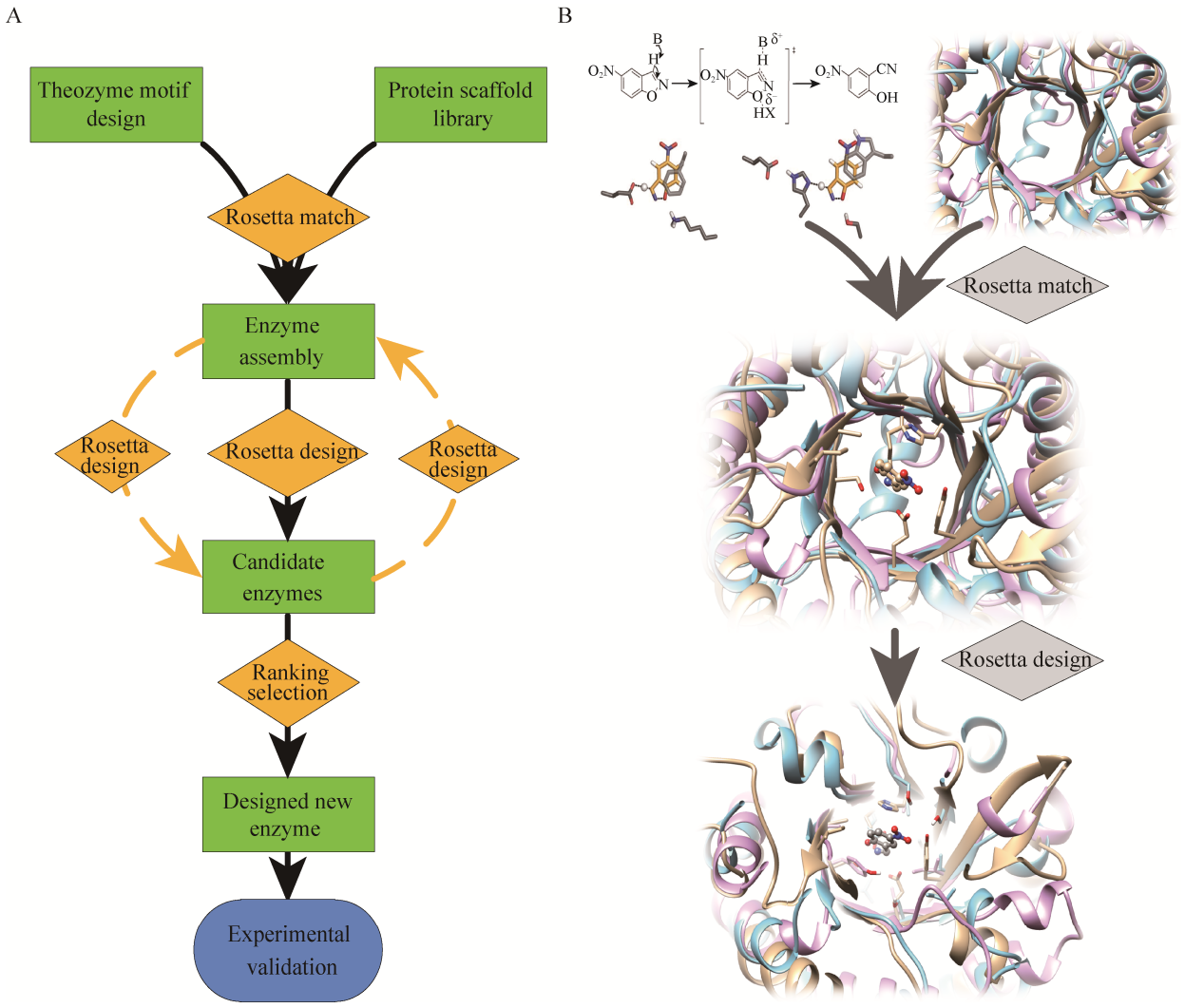

图 1　酶设计的典型流程图 (A)和 kemp 消除反应的酶设计流程 (B)[35]

Fig. 1　Typical flowchart of the enzyme design protocol (A) and the kemp elimination enzyme design flowchart (B). Images of active sites and protein structures were produced in Chimera[35].

表 1　新酶设计的典型案例列表

Table 1　The list of typical examples for new enzyme design

New enzyme reactions	Mainly used software	Authors
Retro-Aldol reaction	Rosetta	Jiang L, et al[22]
Kemp elimination	Rosetta	Röthlisberger D, et al[14]
Diels-Alder reaction	Rosetta	Siegel JB, et al[23]

3　新基因设计的挑战与机遇

尽管新酶设计已经取得了一定的成功，但是依然面临诸多挑战[30]。首先，新基因设计成功率还较低。活性中心的催化基团与过渡态底物模型构建，模型与骨架蛋白匹配，骨架蛋白质的残基构象等都会影响新酶设计的成功率或新酶的性能。例如在 Kemp 消除反应的酶设计中，羧酸基

团作为广义碱与非极性底物间存在相互作用，但是由于羧酸基团的构象自由度较大，如果不能准确计算羧酸基去溶剂化效应的能量消耗及熵减，可能使羧酸基团不适合行使广义碱的作用，进而使反应无法发生[14]。因此我们还需要深入研究酶的催化机理及其计算模拟，如优化分子力场准确计算催化位点与底物、溶剂等作用力[31]，改进催化过渡态能垒计算方法，改善蛋白骨架构象模拟方法等。同时由于蛋白质每个位点都有 20 种可能性，氨基酸之间的相互作用包括氢键、范德华力等多种分子作用力，还存在与溶剂、底物、产物等相互关系，因此蛋白质结构预测计算难度极大，计算设计需要的计算机资源也非常高。Baker 等构建了一套基于 Rosetta 的计算平台(http://boinc.bakerlab.org/rosetta/)，可以通过蛋白质设计爱好者共享计算资源来满足蛋白质结构预测的需求。因此我们还需要通过优化新酶设计算法和计算资源增加新酶设计成功率，为生物催化创造出更多令人惊奇的新反应。

其次，新蛋白酶的设计需要将催化活性中心与蛋白质骨架进行嵌合，而嵌合过程难免会影响蛋白质的结构和稳定性。同时蛋白质骨架与酶催化过程的协同还需要进一步优化，因此新设计的蛋白酶催化活性都普遍偏低，还达不到工业生产的要求。研究人员需要利用经典的酶定向进化和理性改造方法提高新设计酶的催化性能。例如，基于 2012 年 Althoff 等新设计出的 Retro-aldol 酶[32]，2016 年 Obexer 等利用超高通量微流控的方法对该酶进行定向进化[33]，最终得到酶活提升 10^9 的突变体。新酶设计方法与传统酶工程方法结合大大提高了新酶设计的实用性，可进一步开发新酶设计的工业应用潜力。

最后，由于新设计的基因都是自然界不存在的基因，基因功能的测试和鉴定离不开 DNA 合成技术。DNA 合成成本大幅降低，为新基因设计提供了极好的发展机遇。目前 DNA 合成技术发展非常快，不仅合成成本大幅降低，合成通量也大幅提高[34]。比如利用高通量 DNA 芯片合成技术，可以设计与合成各种突变类型的新基因，结合高通量筛选技术，实现新基因合成、密码子优化和酶活改造等多种功能于一体。因此，新酶设计方法与高通量自动化 DNA 合成技术结合，可进一步提高新酶设计的成功率和新酶的表达催化性能，实现新基因的按需设计，满足工业生产需求。

4 结语

合成生物学以工程化理念为导向，对生物体进行有目标的设计、改造乃至重新合成。合成生物学促进了对生命密码从“读识”到“设计”的质变，对揭示生命本质具有重要意义，而由此形成的创新思想、使能技术及工程平台，能促进生物技术革命，被预测为可望改变世界的十大颠覆性技术之一。随着合成生物技术的快速发展，以此为基础的新基因设计技术，将颠覆式创造各种新功能基因，完全突破自然进化的局限，加快新生物功能基因产生速度。按照人类需求快速创建的新功能基因，可极大提高自然生物功能改造与创新的速度，并由此可创建超越自然功能的“人造生物体系”，为解决工业、农业、医药等领域的重大需求提供全新的生物学方案，为我国转变发展方式、引领产业创新发展提供重要技术支撑。

REFERENCES

[1] Ohno S. Evolution by gene duplication. New York:

Springer-Verlag, 1970.
[2] Long M, Langley CH. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. Science, 1993, 260(5104): 91–95.
[3] Long MY, Betrán E, Thornton K, et al. The origin of new genes: glimpses from the young and old. Nat Rev Genet, 2003, 4(11): 865–875.
[4] Zhou Q, Zhang GJ, Zhang Y, et al. On the origin of new genes in *Drosophila*. Genome Res, 2008, 18(9): 1446–1455.
[5] Li D, Dong Y, Jiang Y, et al. A *de novo*-originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res, 2010, 20(4): 408–420.
[6] Li D, Yan ZH, Lu LN, et al. Pleiotropy of the *de novo*-originated gene MDF1. Sci Rep, 2014, 4: 7280.
[7] Chen SD, Zhang YE, Long MY. New genes in *Drosophila* quickly become essential. Science, 2010, 330(6011): 1682–1685.
[8] Carvunis AR, Rolland T, Wapinski I, et al. Proto-genes and *de novo* gene birth. Nature, 2012, 487(7407): 370–374.
[9] Stern-Ginossar N, Weisburd B, Michalski A, et al. Decoding human cytomegalovirus. Science, 2012, 338(6110): 1088–1093.
[10] Liu XQ, Jiang HF, Gu ZL, et al. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. Proc Natl Acad Sci USA, 2013, 110(29): 11928–11933.
[11] Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science, 2000, 290(5494): 1151–1155.
[12] Zhang CT. Advances in synthetic biology studies. Bull Natl Nat Sci Found China, 2009, 23(2): 65–69 (in Chinese).
张春霆. 合成生物学研究的进展. 中国科学基金, 2009, 23(2): 65–69.
[13] Liu D, Du J, Zhao GR, et al. Applications of synthetic biology in medicine and energy. CIESC J, 2011, 62(9): 2391–2397 (in Chinese).
刘夺, 杜瑾, 赵广荣, 等. 合成生物学在医药及能源领域的应用. 化工学报, 2011, 62(9): 2391–2397.
[14] Röthlisberger D, Khersonsky O, Wollacott AM, et al. Kemp elimination catalysts by computational enzyme design. Nature, 2008, 453(7192): 190–195.
[15] Koga N, Tatsumi-Koga R, Liu GH, et al. Principles for designing ideal protein structures. Nature, 2012, 491(7423): 222–227.
[16] Hu XJ, Rousseau R. From a word to a world: the current situation in the interdisciplinary field of synthetic biology. Peer J, 2015, 3(6102): e728.
[17] Wang YH, Wei KY, Smolke CD. Synthetic biology: advancing the design of diverse genetic systems. Ann Rev of Chem Biomol Engin, 2013, 4(1): 69–102.
[18] Gibson DG, Glass JI, Lartigue C, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science, 2010, 329(5987): 52–56.
[19] Regan L, DeGrado WF. Characterization of a helical protein designed from first principles. Science, 1988, 241(4868): 976–978.
[20] Smith BA, Hecht MH. Novel proteins: from fold to function. Curr Opin Chem Biol, 2011, 15(3): 421–426.
[21] Xiong P, Wang M, Zhou XQ, et al. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nat Commun, 2014, 5: 5330.
[22] Jiang L, Althoff EA, Clemente FR, et al. *De novo* computational design of retro-aldol enzymes. Science, 2008, 319(5868): 1387–1391.
[23] Siegel JB, Zanghellini A, Lovick HM, et al. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science, 2010, 329(5989): 309–313.
[24] Richter F, Leaver-Fay A, Khare SD, et al. *De novo* enzyme design using Rosetta3. PLoS ONE, 2011, 6(5): e19230.
[25] Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res, 2000, 28(1): 235–242.
[26] Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res, 2002, 30(1): 47–49.
[27] Zanghellini A, Jiang L, Wollacott AM, et al. New algorithms and an in silico benchmark for

computational enzyme design. Protein Sci, 2006, 15(12): 2785–2794.
[28] Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA, 2000, 97(19): 10383–10388.
[29] Siegel JB, Smith AL, Poust S, et al. Computational protein design enables a novel one-carbon assimilation pathway. Proc Natl Acad Sci USA, 2015, 112(12): 3704–3709.
[30] Huang PS, Boyken SE, Baker D. The coming of age of *de novo* protein design. Nature, 2016, 537(7620): 320–327.
[31] Ponder JW, Case DA. Force fields for protein simulations. Adv Protein Chem, 2003, 66: 27–85.
[32] Althoff EA, Wang L, Jiang L, et al. Robust design and optimization of retroaldol enzymes. Protein Sci, 2012, 21(5): 717–726.
[33] Obexer R, Godina A, Garrabou X, et al. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. Nat Chem, 2016, doi: 10.1038/nchem.2596.
[34] Kosuri S, Church GM. Large-scale *de novo* DNA synthesis: technologies and applications. Nat Methods, 2014, 11(5): 499–507.
[35] Pettersen EF, Goddard TD, Huang CC, et al. UCSF chimera—a visualization system for exploratory research and analysis. J Comput Chem, 2004, 25(13): 1605–1612.

(本文责编　陈宏宇)

江会锋　博士，中国科学院天津工业生物技术研究所研究员，博士生导师，中国科学院“百人计划”，天津市“青年千人计划”，天津市创新人才推进计划入选者。2015 年任中科院系统微生物工程重点实验室副主任。主攻方向为新酶设计和酵母基因组工程。已先后在 *Proc Natl Acad Sci USA*、*Genome Res*、*PLoS Genetics*、*Cell Res* 等国际学术期刊上发表 SCI 论文 20 余篇，影响因子总计超过 120，他引总计 400 余次，申请国家专利 5 项。承担国家 973 计划项目子课题 1 项；国家 863 计划项目子课题 1 项；国家自然科学基金项目 3 项；天津市科技支撑计划项目 2 项；企业联合研发项目 3 项。