

可理解组合语义的大肠杆菌重编程方法

陈 梅

中央民族大学信息工程学院，北京

收稿日期：2025年10月5日；录用日期：2025年10月29日；发布日期：2025年11月6日

摘 要

理解组合语义在人类交流中至关重要。本研究开发出一种能够编程活细胞的策略，利用CRISPR/Cas9系统的精准基因组编辑能力来解析组合语义。在我们的实验中，大肠杆菌成功区分了“good morning”与“morning good”。这项研究不仅证明基因组编辑技术可用于存储或记录信息，更能用于识别信息。通过该策略，人类与活细胞的通信范围将拓展至任何可编码为DNA的信息，这将推动智能遗传器件的构建。我们相信该策略在生物计算、生物传感、生物治疗等领域具有广阔的应用前景。

关键词

DNA计算，DNA存储，基因电路，CRISPR/Cas9

Programming *Escherichia coli* to Understand Compositional Syntax

Mei Chen

School of Information Engineering, Minzu University of China, Beijing

Received: October 5, 2025; accepted: October 29, 2025; published: November 6, 2025

Abstract

Understanding compositional syntax is important in human communication. Here, we developed a strategy which can program living cells to understand compositional syntax through precise genome editing ability of CRISPR/Cas9 system. In our example, the *E. coli* successfully tells the difference between “good morning” and “morning good”. This study shows not only can genome editing technology be used in storing or recording information, but it can also be used in understanding information. Through this strategy, the range that people can communicate with living cells will widen to any information which can be encoded into DNA. This will advance the construction of intelligent genetic devices. We believe that this strategy has many potential applications in biocomputing,

biosensing, biotherapy, and so on.

Keywords

DNA Computing, DNA Storage, Genetic Circuit, CRISPR/Cas9

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着 DNA 技术的飞速发展,近年来已成功构建出多种适用于体外和体内的 DNA 电路。这些 DNA 电路可广泛应用于逻辑计算[1]-[3]、数据存储[4] [5]、并行计算[6]、传感器[7]及智能计算[8] [9]等领域。

在基因组中留下永久标记是实现 DNA 电路记忆功能的传统方法[10]。目前已有多种技术被应用于构建记忆装置,包括 DNA 重组酶[11]-[14]、单链 DNA 重组工程[15]以及成簇规律间隔短回文重复序列/CRISPR 关联蛋白 9 系统(CRISPR/Cas9 系统) [5] [16] [17]。CRISPR/Cas9 系统作为高效的基因组工程工具近年来备受关注[18]。该系统通过 sgRNA 引导核酸酶 Cas9 蛋白至特定 DNA 序列(原间隔区),其 3'末端需存在原间隔区相邻基序(PAM, 序列为 5'-NGG-3')。sgRNA 由引导序列和支架序列构成:引导序列通常为 20 bp,可根据实际需求人工设计以指定靶向 DNA 位点;支架序列为 80 bp,负责介导 sgRNA 与 Cas9 的结合。当 Cas9 催化产生 DNA 双链断裂后,细胞会通过易出错的非同源末端连接或同源重组修复机制进行修复(图 1)。

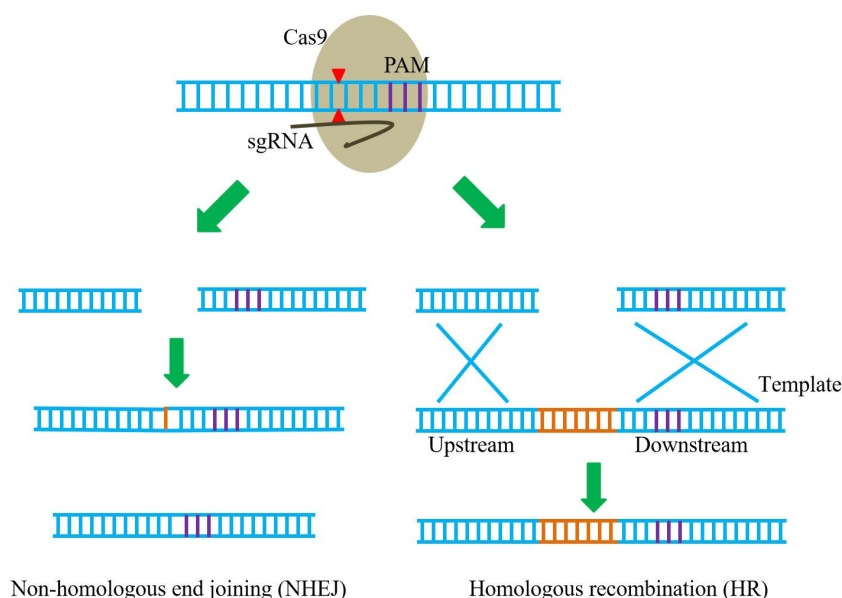


Figure 1. The pathway for genome editing by CRISPR/Cas9. Left: NHEJ, the double-stranded break are repaired by endogenous DNA repair machinery and rejoined, which can lead to random indel mutations resulting in gene knockout. Right: HR, which allows precise editing according to the template

图 1. CRISPR/Cas9 系统进行基因组编辑的路径。左侧: 非同源末端连接修复路径。双链断裂由细胞内源 DNA 修复机制进行修复并重新连接, 此过程可能产生随机的插入或缺失突变, 从而导致基因敲除。右侧: 同源重组修复路径。该机制能够根据提供的修复模板实现精确的基因编辑

人类语言能够通过组合不同词汇形成组合语义，从而用有限词汇表达无限含义。理解组合语义这种基于记忆的复杂智能活动，可确保沟通的有效性。基于 CRISPR/Cas9 基因组编辑实现的记忆功能，本研究成功编程大肠杆菌使其能够识别“good morning”这一组合语义。当输入语序颠倒的“morning good”时，大肠杆菌则不会产生响应。

对错误语序，则按相反顺序输入。通过菌落 PCR 检测结果，引物根据识别起始位点上下游序列设计。正确语序“good morning”的结果比错误语序长约 1 kb，这是因为“morning”词汇携带 1 kb 特征编码 CC2 (图 3)。测序结果证实大肠杆菌成功识别了“good morning”的输入信息。

2. 系统设计

该系统在基因组中定义一个识别起始位点来读取输入信息。若对细胞非必需，外显子或内含子均可被选作识别起始位点。词汇被分别编码为 DNA 序列，每个词汇具有唯一的特征编码用于识别。研究构建了三种 DNA 电路来识别组合语义(图 2)。

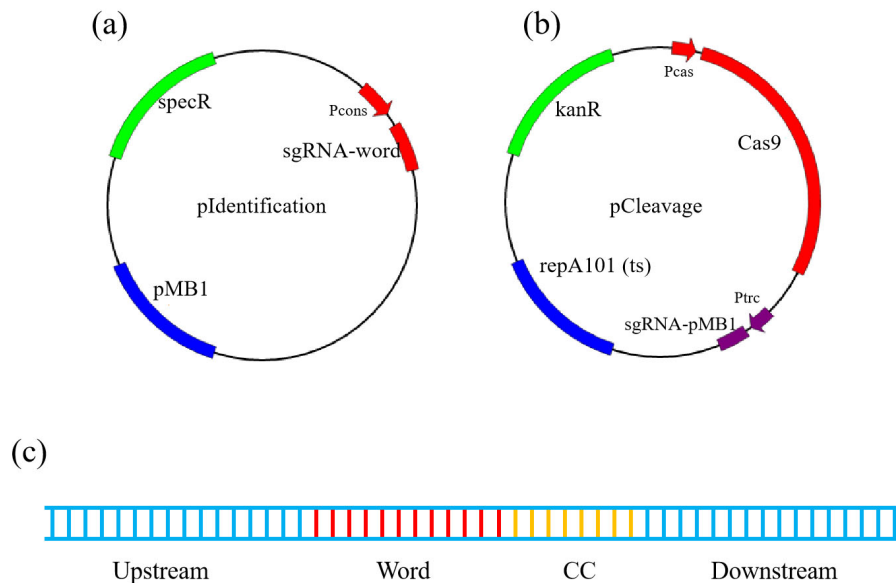


Figure 2. (a) Identification circuit. (b) Cleavage circuit. (c) Word template circuit
图 2. (a) 识别电路。 (b) 切割电路。 (c) 词汇模板电路

第一种是识别电路，为质粒电路，用于识别词汇序列。组成型启动子 *Pcons* 驱动靶向末位词汇特征编码的 *sgRNA* 转录，该电路使用 *pMB1* 复制子。第二种是切割电路，同为质粒电路，用于提供 *Cas9* 蛋白。天然启动子 *Pcas* 驱动 *Cas9* 表达，IPTG 诱导型启动子 *Ptrc* 则驱动靶向识别电路 *pMB1* 的 *sgRNA* 转录，该电路使用温度敏感型复制子 *repA101(Ts)*。第三种是词汇模板电路，为双链 DNA 结构，包含上游序列、词汇、特征编码和下游序列。每个词汇均设计有专属特征编码、识别电路和词汇模板，CRISPR/Cas9 系统负责准确读取词汇信息，最终结果可通过末位词汇的特征编码进行检测。

在读取输入时，首先将切割电路转入细胞以提供 *Cas9* 蛋白。读取单个词汇时，将其识别电路与词汇模板电路共转化至细胞内。识别电路中的 *sgRNA* 会靶向末位词汇的特征编码，从而引导 *Cas9* 蛋白精确定位，并根据词汇模板电路引入目标词汇。

完成词汇读取后，识别电路与词汇模板电路会被清除以进行下一轮读取。识别电路通过 IPTG 诱导清除：IPTG 激活切割电路中 *Ptrc* 启动子转录靶向 *pMB1* 的 *sgRNA*，在 *sgRNA* 与 *Cas9* 作用下，*pMB1* 通

过非同源末端连接机制被破坏,从而清除以 pMB1 为复制子的识别电路。词汇模板电路因无法自我复制的双链 DNA 特性,在培养过程中可自然清除。完成所有词汇读取后,通过升温诱导切割电路自我清除,该过程利用其温度敏感型复制子 repA101(Ts)的特性。

本实验中,首先在基因组中设定识别起始位点。选择大肠杆菌非必需基因 *aroA* 作为识别起始位点,为确保可靶向性,在该位点选取邻近 PAM 序列的 20 bp 特征编码 CC0。其次构建词汇模板:将“good”编码为 DNA 序列并设计 23 bp 且 3'端含 PAM 的特征编码 CC1,将 CC0 上游、“good”DNA 序列、CC1 及 CC0 下游连接构成“good”词汇模板;将“morning”编码为 DNA 序列,因其为末位词汇,特意设计约 1 kb 的长特征编码 CC2 便于检测,将 CC1 上游、“morning”DNA 序列、CC2 及 CC1 下游连接构成“morning”词汇模板。

随后为每个词汇设计识别电路:针对“good”基于 CC0 设计 sgRNA 构建 pIdentification-good 质粒;针对“morning”基于 CC1 设计 sgRNA 构建 pIdentification-morning 质粒。最后分别将“good morning”与“morning good”输入大肠杆菌 DH5 α : 对于正确语序,同步输入 pIdentification-good 与“good”模板后,再同步输入 pIdentification-morning 与“morning”模板;对于错误语序,则按相反顺序输入。通过菌落 PCR 检测结果,引物根据识别起始位点上下游序列设计。正确语序“good morning”的结果比错误语序长约 1 kb,这是因为“morning”词汇携带 1 kb 特征编码 CC2 (图 3)。测序结果证实大肠杆菌成功理解了“good morning”的输入信息。

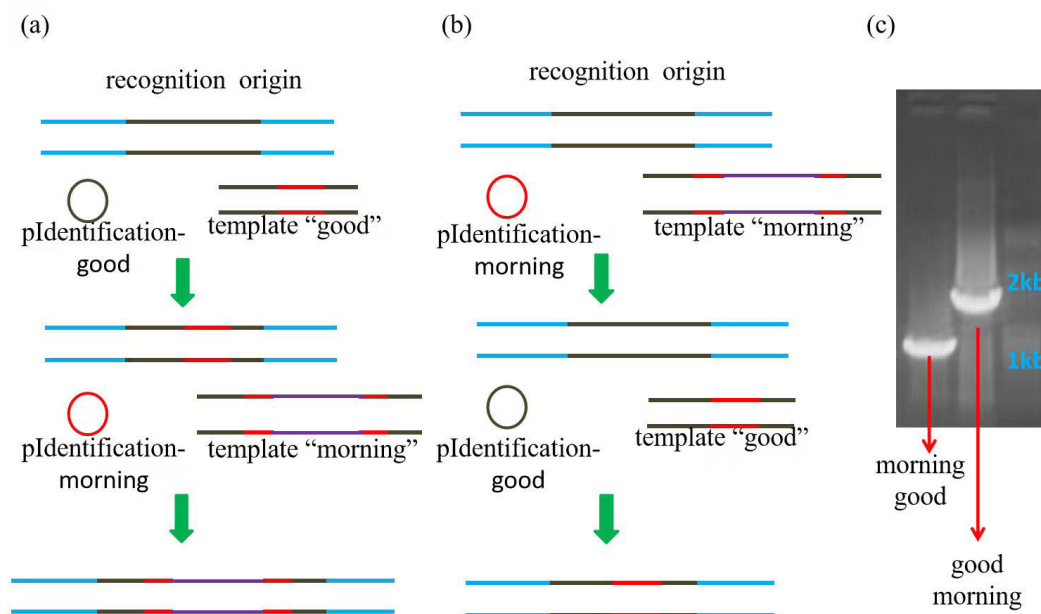


Figure 3. The reading procedure of recognition origin. (a) The input is “good morning”. (b) The input is “morning good”. (c) Results of colony PCR

图 3. 识别起始位点的读取流程。(a) 输入为“good morning”的情况。(b) 输入为“morning good”的情况。(c) 菌落 PCR 结果

3. 材料和方法

3.1. 质粒和寡核苷酸

本研究使用的质粒与寡核苷酸详见表 1 和表 2。所有寡核苷酸均由 ToloBio 合成。pIdentification-good 质粒的构建方法为:使用引物 pIdentification-good-F 和 pTargetF-R 对 pTargetF 质粒进行扩增。

pIdentification-morning 质粒的构建则采用引物 pIdentification-morning-F 和 pTargetF-R。PCR 产物经 DpnI 内切酶处理，再通过 T4 DNA 连接酶和 T4 多聚核苷酸激酶进行自连。最终将所得质粒转化至大肠杆菌 DH10B 感受态细胞，并通过测序验证构建结果。

Table 1. Plasmids used in this study
表 1. 本研究使用的质粒列表

Plasmid	Characteristics	Source
pCleavage	It contains a Cas9 gene with a Pcas promoter, an isopropyl β -D-1-thiogalactopyranoside (IPTG)-inducible sgRNA, which can guide Cas9 to the pMB1 replicon of pTargetF, and a temperature-sensitive replication repA101(Ts).	[19] (Addgene 62225)
pTargetF	It is used to express the targeting sgRNA.	[19] (Addgene 62226)
pIdentification-good	It is used to express the sgRNA which targets the recognition origin.	This study
pIdentification-good	It is used to express the sgRNA which targets the word “good”.	This study

Table 2. Oligonucleotides used in this study
表 2. 本研究使用的寡核苷酸

Oligonucleotide	Sequence (5'-3')
pIdentification-good-F	GCAGCAGCAGCAATCAAAGGGTTTTAGAGCTAGAAATAGCAAG
pIdentification-morning-F	AGGTATAATACTAGTCAATCGTACGTACGTGCGACGTTTTAGAGCTAGAAATAGC
pTargetF-R	ACTAGTATTATACCTAGGAC
UP-F	GGTCCATTACACGCAGAAGG
Good-R	CGTACGATTGTGCGAACGAACGAGTGTGCTGCTGCCAGAAAGTAAG
Good-F	CGTTCGCACAATCGTACGTACGTGCGACGGGAAAGTGACCGGTATTGGACG
DN-R	CAACGCCACCAGCGAGAAAC
Morning-F	GTTCTAGCGTGCGGCCGTGCGCTGACGGGAAAGTGACCGGTATTG
Morning-R	GCACGTACGTACGATTGTGC
MorningCC2-F	GCACAATCGTACGTACGTGCAAGGCCAGTCTTTCGACTG
MorningCC2-R	CGCACGGCCGCACGCTAGAACGGACGCTCGAGTCCCTATCAGTGATAG

3.2. 编码

单词 “good” 根据 ASCII 编码转换为二进制字符串 “01000111011011110110111101100100”，随后按照规则(00-A, 01-C, 10-G, 11-T)转换为 DNA 编码 “CACTCGTTCGTTTCGCA”。为此词汇设计了特征编码 CC1 “CAATCGTACGTACGTGCGACGGG”。在构建 “good” 模板时，提取大肠杆菌 DH5 α 基因组，使用引物 UP-F 与 Good-R 扩增获得 522 bp 的上游片段，使用引物 Good-F 与 DN-R 扩增获得 408 bp 的下游片段。最后以 UP-F 和 DN-R 为引物对上下游片段进行融合扩增，获得完整的 “good” 模板。

单词 “morning” 根据 ASCII 编码转换为二进制字符串 “01101101011011110111001001101110011010010110111001100111”，并沿用相同规则转换为 DNA 编码 “CGTCCGTTCTAGCGTGCGGCCGTGCGCT”。为其设计的特征编码 CC2 序列为 (CTCGAGTCCCTATCAGTGATAGAGATTGACATCCCTATCAGTGATAGAGATACTGAG-CACATCAGCAGGACGCACTGACCGAATTCAACAATAGCATACATTATACGAAGTTATAAAGAGGA

GAAAGGTACCATGGTGAGCAAGGGCGAGGAGCTGTTCAC-
CGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCG
GCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGG-
CAAGCTGCCCCGTGCCCTGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAATGCTTCGCCCC
CTACCCCGACCACATGAAGCTGCACGACTTCTTCAAGTCCGCCATGCCCCGAAGGCTAC-
GTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTT
CGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGG-
CAACATCCTGGGGCACAAGCTGGAGTACAACAGCCACAACGTCTATATCATGGCCGACA
AGCAGAAGAACGGCATCAAGGTGAAGTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTG-
CAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCCGACAA
CCACTACCTGAGCTACCAGTCCGCCCTGAGCAAAGACCCCCAACGAGAA-
GCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCT
GTACAAGTAATAAAAGCTTAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCAC-
CTGACCCCATGCCGAAGTCAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCAT
GCGAGAGTAGGGAAGTCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTT)。

本研究由 ToloBio 合成 morning-CC2 片段。使用引物 UP-F 与 Morning-R 获得 555 bp 上游片段, Morning-F 与 DN-R 获得 437 bp 下游片段, MorningCC2-F 与 MorningCC2-R 获得 1074 bp 插入片段。最终通过引物 UP-F 和 DN-R 对上游片段、插入片段及下游片段进行融合扩增, 构建出完整的“morning”模板。

3.3. 阅读

采用标准热激法将 pCleavage 质粒转化至细胞内, 并添加 10 mM 阿拉伯糖诱导 λ -Red 重组系统表达。读取“good”时, 将 pIdentification-good 质粒与“good”模板共转化至细胞; 读取“morning”时, 将 pIdentification-morning 质粒与“morning”模板共转化。每个词汇读取完成后, 使用 0.5 mM IPTG 诱导过夜以清除识别电路。电转参数设置为 2.5 kV、200 Ω , 转化后细胞在 30℃ 复苏 1 小时, 再涂布于含卡那霉素(50 mg/L)和壮观霉素(50 mg/L)的 LB 琼脂平板, 30℃ 培养过夜。通过 DNA 测序验证转化子。最终, 将菌落置于 37℃ 培养过夜以实现 pCleavage 质粒的自我清除。

4. 结论

在本研究中, 我们基于 CRISPR/Cas9 系统精准基因组编辑实现的记忆功能, 构建了一套可编程活细胞识别组合语义的系统。为确保基因组编辑的精确性, 需要避免 CRISPR/Cas9 系统的脱靶效应干扰特异性识别: 首先, 每个词汇识别电路中的 sgRNA 必须严格遵循 CRISPR/Cas9 系统标准进行设计; 其次, 采用噬菌体源重组酶系统辅助实现精准基因组编辑[19]。

在精准基因组编辑基础上, 系统通过巧妙设计实现组合语义识别: 一方面, 为每个词汇设计专用识别电路, 确保词汇仅能在正确序列中被读取; 另一方面, 系统中所有电路均可被清除, 从而实现多词汇顺序读取。单个词汇读取后, 其识别电路与词汇模板即被清除; 完成所有词汇读取后, 切割电路也被清除。

识别组合语义是学习、判断、推理与控制等智能活动的基础。本研究展现了利用可编程基因组编辑技术构建智能遗传器件的巨大潜力。当前文本、图像、影像等各类信息均可存储于 DNA 中[5] [20] [21], 这使得人类与活细胞的通信范围可拓展至任何可编码为 DNA 的信息。此外, 该系统有利于细胞行为调控: 基于此系统, 由于细胞能根据不同输入序列作出差异化响应, 未来可能构建复杂的时序逻辑器件。

因此, 识别组合语义将有效推进智能遗传器件的构建。我们相信这项研究在生物计算、生物传感及生物治疗等领域具有广阔应用前景。

当然, 当前系统仍存在一定局限性。首先, 基因组编辑效率较低, 识别过程较慢; 其次, 系统的可扩展性仍显不足, 随着词汇数量的增加, 细胞资源问题将愈发显著; 此外, 系统在复杂生理环境下的稳定性和鲁棒性也有待进一步提高。未来研究可从优化基因编辑技术、构建分布式系统等方面着手, 构建出更高效、更具扩展性的智能生物计算系统。

致谢

本研究得到中央民族大学“有组织科研项目”资助。

参考文献

- [1] Tamsir, A., Tabor, J.J. and Voigt, C.A. (2010) Robust Multicellular Computing Using Genetically Encoded NOR Gates and Chemical ‘Wires’. *Nature*, **469**, 212-215. <https://doi.org/10.1038/nature09565>
- [2] Siuti, P., Yazbek, J. and Lu, T.K. (2013) Synthetic Circuits Integrating Logic and Memory in Living Cells. *Nature Biotechnology*, **31**, 448-452. <https://doi.org/10.1038/nbt.2510>
- [3] Beiki, Z. and Jahanian, A. (2017) DENA: A Configurable Microarchitecture and Design Flow for Biomedical DNA-Based Logic Design. *IEEE Transactions on Biomedical Circuits and Systems*, **11**, 1077-1086. <https://doi.org/10.1109/tbcas.2017.2708747>
- [4] Erlich, Y. and Zielinski, D. (2017) DNA Fountain Enables a Robust and Efficient Storage Architecture. *Science*, **355**, 950-954. <https://doi.org/10.1126/science.aaj2038>
- [5] Shipman, S.L., Nivala, J., Macklis, J.D. and Church, G.M. (2017) CRISPR-Cas Encoding of a Digital Movie into the Genomes of a Population of Living Bacteria. *Nature*, **547**, 345-349. <https://doi.org/10.1038/nature23017>
- [6] Xu, J. (2016) Probe Machine. *IEEE Transactions on Neural Networks and Learning Systems*, **27**, 1405-1416. <https://doi.org/10.1109/tnnls.2016.2555845>
- [7] Hsu, C., Chen, B., Hu, R. and Chen, B. (2016) Systematic Design of a Quorum Sensing-Based Biosensor for Enhanced Detection of Metal Ion in *Escherichia coli*. *IEEE Transactions on Biomedical Circuits and Systems*, **10**, 593-601. <https://doi.org/10.1109/tbcas.2015.2495151>
- [8] Chen, M. and Xu, J. (2015) Construction of a Genetic Conditional Learning System in *Escherichia coli*. *Science China Information Sciences*, **58**, 1-6. <https://doi.org/10.1007/s11432-015-5308-8>
- [9] George, A.K. and Singh, H. (2017) DNA Implementation of Fuzzy Inference Engine: Towards DNA Decision-Making Systems. *IEEE Transactions on NanoBioscience*, **16**, 773-782. <https://doi.org/10.1109/tnb.2017.2760821>
- [10] Burrill, D.R. and Silver, P.A. (2010) Making Cellular Memories. *Cell*, **140**, 13-18. <https://doi.org/10.1016/j.cell.2009.12.034>
- [11] Friedland, A.E., Lu, T.K., Wang, X., Shi, D., Church, G. and Collins, J.J. (2009) Synthetic Gene Networks That Count. *Science*, **324**, 1199-1202. <https://doi.org/10.1126/science.1172005>
- [12] Yang, L., Nielsen, A.A.K., Fernandez-Rodriguez, J., McClune, C.J., Laub, M.T., Lu, T.K., *et al.* (2014) Permanent Genetic Memory with > 1-Byte Capacity. *Nature Methods*, **11**, 1261-1266. <https://doi.org/10.1038/nmeth.3147>
- [13] Bonnet, J., Yin, P., Ortiz, M.E., Subsoontorn, P. and Endy, D. (2013) Amplifying Genetic Logic Gates. *Science*, **340**, 599-603. <https://doi.org/10.1126/science.1232758>
- [14] Roquet, N., Soleimany, A.P., Ferris, A.C., Aaronson, S. and Lu, T.K. (2016) Synthetic Recombinase-Based State Machines in Living Cells. *Science*, **353**, aad8559. <https://doi.org/10.1126/science.aad8559>
- [15] Farzadfard, F. and Lu, T.K. (2014) Genomically Encoded Analog Memory with Precise *in Vivo* DNA Writing in Living Cell Populations. *Science*, **346**, Article ID: 1256272. <https://doi.org/10.1126/science.1256272>
- [16] Perli, S.D., Cui, C.H. and Lu, T.K. (2016) Continuous Genetic Recording with Self-Targeting CRISPR-Cas in Human Cells. *Science*, **353**, aag0511. <https://doi.org/10.1126/science.aag0511>
- [17] Sheth, R.U., Yim, S.S., Wu, F.L. and Wang, H.H. (2017) Multiplex Recording of Cellular Events over Time on CRISPR Biological Tape. *Science*, **358**, 1457-1461. <https://doi.org/10.1126/science.aao0958>
- [18] Doudna, J.A. and Charpentier, E. (2014) The New Frontier of Genome Engineering with CRISPR-Cas9. *Science*, **346**, Article ID: 1258096. <https://doi.org/10.1126/science.1258096>
- [19] Jiang, Y., Chen, B., Duan, C., Sun, B., Yang, J. and Yang, S. (2015) Multigene Editing in the *Escherichia coli* Genome

- via the CRISPR-Cas9 System. *Applied and Environmental Microbiology*, **81**, 2506-2514.
<https://doi.org/10.1128/aem.04023-14>
- [20] Church, G.M., Gao, Y. and Kosuri, S. (2012) Next-Generation Digital Information Storage in DNA. *Science*, **337**, 1628-1628. <https://doi.org/10.1126/science.1226355>
- [21] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B., *et al.* (2013) Towards Practical, High-Capacity, Low-Maintenance Information Storage in Synthesized DNA. *Nature*, **494**, 77-80.
<https://doi.org/10.1038/nature11875>