Embedding DNA-Based Natural Language in Microbes for the Benefit of Future Researchers

Heqian Zhang[1], Jiaquan Huang[1], Xiaoyu Wang[1], Zhizeng Gao[2], Song Meng[3], Hang Li[4], Shanshan Zhou[5], Shang Wang[6], Shan Wang[7], Xunyou Yan[8], Xinwei Yang[9], Xiaoluo Huang[10*], Zhiwei Qin[1*]

[1]Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong, 519087, China. [2]School of Marine Sciences, Sun Yat-sen University, Zhuhai, Guangdong, 519000, China. Southern Laboratory of Ocean Science and Engineering, Zhuhai, Guangdong, 519000, China. [3]State Key Laboratory of Drug Research & Natural Products Research Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China. University of Chinese Academy of Sciences, Beijing, 100049, China. Zhongshan Institute for Drug Discovery, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Zhongshan, Guangdong, 528400, China. [4]School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou, Guangdong, 510006, China. [5]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, 100101, China. [6]College of Plant Science, Jilin University, Changchun, Jilin, 130062, China. [7]State Key Laboratory of Microbial Technology, Shandong University, Qingdao, Shandong, 266237, China. [8]College of Life Science, Langfang Normal University, Langfang, Hebei, 065000, China. [9]School of Pharmaceutical Sciences, Zhengzhou University, Zhengzhou, Henan, 450001, China. [10]Shenzhen Key Laboratory of Synthetic Genomics, Guangdong Provincial Key Laboratory of Synthetic Genomics, Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, 518055, China.

Correspondence: Zhiwei Qin, Xiaoluo Huang
E-mail: z.qin@bnu.edu.cn; huangxl@siat.ac.cn

Material and methods

Encoding the text data in a single DNA strand

The algorithm is accessible on the Storage-D platform (http://storage.dailab.xyz:16666/). During the encoding process, two 900 base pair (bp) fragments were generated. To ensure synthesis feasibility and sequence stability, homopolymer stretches were limited to fewer than six nucleotides, and the GC content was targeted to range between 40% and 60%. Each 900 bp fragment was augmented with a 2-byte Reed-Solomon (RS) error correction code, following the platform's protocols. These fragments were then ligated to form a contiguous 1.8 kilobase (kb) sequence. To facilitate specific amplification, a set of 20-base pair random access primers was appended to the concatenated sequence, designed using the primer design tool integrated into the Storage-D platform.

General experiment for bacterial transformation

The DNA sequence was synthesized by GENEWIZ (Suzhou, China) and subsequently cloned into the *Streptomyces* integrative plasmid pIB139 via homologous recombination at the NdeI and EcoRI restriction sites. This plasmid was then transformed into *E. coli* ET12567/pUZ8002 and cultured on Luria-Bertani (LB) agar plates containing apramycin (25 μg/mL), kanamycin (50 μg/mL), and chloramphenicol (25 μg/mL) at 37 °C overnight. Colonies were isolated and verified through sequencing.

The PCR reaction was set up in a 50 μL system containing 25 μL 2× Phanta Max Buffer, 1 μL dNTP, 1 μL DNA Polymerase, 1 μL each of forward and reverse primers, 1 μL template DNA, and 20 μL ddH2O. The PCR conditions were: initial denaturation at 95 °C for 3 minutes; 35 cycles of 98 °C for 10 seconds, 60 °C for 30 seconds, and 72 °C for 2 minutes; followed by a final extension at 72 °C for 7 minutes.

The confirmed colony was grown in fresh LB liquid medium with antibiotics under constant shaking (200 rpm) until the culture reached the exponential phase (OD_{600} ~0.8). The culture broth was then harvested by centrifugation at 5,000 rpm for 10 minutes and washed three times with LB medium without antibiotics.

To obtain *Streptomyces* spores, strain G7 was cultured on MS solid medium (20 g mannitol, 20 g soya flour, 20 g agar, and 1 L tap water) at 30 °C for 15 days. Spores were collected by gently washing the surface of the strain with 20% glycerol using a sterile cotton bud. The spores were washed three times with tryptic soy broth (TSB)

medium at 12,000 rpm for 5 minutes to remove glycerol, then resuspended in TSB and incubated at 50 °C for 10 minutes before cooling to room temperature.

Equal volumes of *E. coli* and G7 spores were mixed, centrifuged at 12,000 rpm for 5 minutes to remove the supernatant, resuspended in 200 μL of TSB, and plated onto MS agar supplemented with 10 mM Mg^{2+}. After incubation at 30 °C for 20 hours, the plates were overlaid with apramycin (25 μg/mL) and nalidixic acid (25 μg/mL) and cultured for 5 days at 30 °C. Colonies were isolated and purified on MS agar with antibiotics for three cycles over 12 days at 30 °C. Finally, the colonies were cultured on MS agar without antibiotics for 5 days at 30 °C and identified by Sanger sequencing.

The information of natural language and its encoded DNA sequence

This is *Streptomyces ginsengnesis* G7 that was discovered in 2020 in Jilin Province China and demonstrated the ability to produce natural products such as lydicamycins which inhibit weed growth, and other structurally unknown compounds which inhibit the growth of fungal pathogens in crops.

GTGAACGCATGGTCACCTACCGTATCAATGCCAAATTTCCAAGCGGCGAAGACCCGGATTGGCCAGA
CCGCGAGGGGTAAGACACGCCCTCGCCCGAAGACTCAAAAACCCCAGCACACGAGAATTCCAAGAC
TCCAAACTACGAGACTGCGAACTGCCGAACCGTAAAACACACGAAGAGACGGACAGGCCAGCCTAC
GAGACGGCGAGATAACCAAAAGTCGAATACAAAAAATTACCAGCATACGAGATAACCAATACTCGA
ACGCCCGAGACGGCGAAAAACCGACCCCGCCCTCAAGAAGAAATACGACCAGCCCAGCGTGCCAAA
TGCCCAATTCGAAAAAACACCGACAGGCCAGAAAGAAAACACGCGAAGACCCGACTTGGCCAACCC
ACCAAGCGGCGAAAAACCGAACAGTCGAAAGATCGAGGGATCCAACACTCGAAGGGTCCAATATCC
GAGCACTCAGACACGCCAGACGCCGAAGCTACCGATTGGCCGAAAACCCGAGGCGCAGAAACCCCG
AGGGGCCGATATTCCAAGCCGCGAAGCATCCGACATTCCTCGCGTCCAGACCGAAAAATCACCAGAC
GCCGAAACATCCGACCGGCCTCCGCGCAAAAACCCCAAACAGCGAGAAAACCAAAGGCCGAGACAG
CCAGCCCCCGAATTCAAAGATACGCATCCACCCGAAATGGCCAAAGTTCGAAACTACAAACACGCGG
ATAGCCCAGCCCGCGAAGCTACCAACCTGCCACCATTCAAACACACCAGCACTCGAGACCACGAGAT
AACGAGACATCCAAAAACCGAGCGCACCAGCGGTCCAGCATTCGAAAATGCCAATACACGAATACA
CCACTTGGCAAAAACACCAGACGGCCAATATCCGGATTGCCCAGAATACAAACAGTTAGTAGGGGA
TAAGGTAGTCAGGATGGATGGTAGCCGTTAAACCAGCGCCTTTGGGTATTGAATTTTTAGGACTCCT
GGTAGGATAGAGGGAAGGGTACATGCGGGCAAGGAACGAACCCCATATGTTTAAATGGTCTTATGG
TATTTGACCCCTCTGGTATTAGTTACACCAGGGATAATGGGTACCTGGGATCCAGGTAATTTGCCGTC
AGAGTATAAAGTTTTTAGTCGATCCAGGATCCCTGGTCTCCTGCGTTGAGTCCAACATGCAAGTTAG
GACCCGGAATGAAATGTCTAAAAGTTTCTAAGGTATTAACCCAACGGAGTGATCAGTATACCACCGA
ACGGACCTACGTGGAGCGATGTGGACCAGCCAGTCAGCCCTGATGGTGGCAGTCCCACGTGGAAAC
ATGCGCATCAGGATTCCTGGTCACGAGGTTGTTTGCGTAGATATATTTTAGGGTATATGGAGTAAAC
TAAACCAGATAATCAGATTACGCCTAAACCAGTACCGAAGAAGTCAAGGTTGTTTGGACTCAAGAAC

CTCACGGTGCGACCGCACGAACCATCAATCGCATACTCCCACAATTAATCCATCGCTCGCACCAACCA
TCCCACGCTCGAACCATCCAATACTCGCACGATCGATCCCTCCAACAAATACTCGCACGCACCCTCCA
TCCATCCAACACTCCCATACACCCACGCTTACTTAATCGATCGCTCCCTTAATCCATTAATCACACACA
CCCACAATCGCACCCTCGCACACTCGATCCCTCGCACAATCCAACACTCGCTCAATCGCTCGATCGCA
CACTCCAATAATCCCTCACTCCCTCAATCAAACCCACCAACAATCGCACCATCCCTCGCTTAAATAATC
GCTCGATGAATCTCATTCCGTCTCCCCCGTCTAAC

NOTE. Sequences highlighted in green are designated for PCR amplification.