DNA Digital Data Storage

(from William and Mary iGEM; contact: igem@wm.edu or mssaha@wm.edu)

Summary

A major problem of this era is to store increasing amounts of data generated worldwide, which is predicted to reach 40 ziga-bytes (ZB) by the year 2020. In the near future, traditional storage devices (such as SSDs, flash devices and SD cards) may not be able to meet our requirements for storing sufficient information. With a perspective of using smaller mediums rather than bigger data storage centers, the key to solving this issue can be found through the use of the DNA molecule. Billions of years of evolution have optimized DNA for transmission and storage of data in a way that is both efficient and reliable. Theoretically, all 40 ZB of man-made data can be stored in just 90 grams of DNA and it could potentially last for hundreds of thousands of years.

Initial attempts to use DNA as an archival medium, dating back to the 1960s, could only go so far because of the difficulties faced when attempting to synthesize long stretches of DNA. However, with the current technological advancements in DNA synthesis and sequencing, synthetic biologists are fast making progress in storing large datasets in DNA. Although current limitations, such as lack of efficient random access read and rewritability features, is keeping this novel solution at the experimental level, the multitude of advantages the DNA-based storage offers is foreseen to reshape the future of communication and data storage in the near future. *{from Genscript}*

In this activity, we will go through the workflow of storing data in DNA.

1. Converting classical sources of information like images or videos into specialized, compressed codes that can be represented by four alphabet letters,
2. Code words are synthesized as DNA molecules and stored,
3. If necessary, written and stored data can be re-written using DNA editing methods,
4. To read the stored data, Sanger sequencing or high-throughput sequencing to access short or large portions of archived data, respectively.

We will also introduce the advantage of applying synthetic biology strategies to store data in DNA.

1. A storage capacity 1,000,000 folds higher than current platforms,
2. Reproducibility through fast and cost-effective PCR,
3. Stability and resistance to change,
4. Passive and cost-efficient maintenance,
5. A hack-proof solution due to lack of remote access,
6. Long lifespan across thousands of years,
7. Relief from technological obsolescence.

As well as the disadvantages of DNA Digital Data Storage:

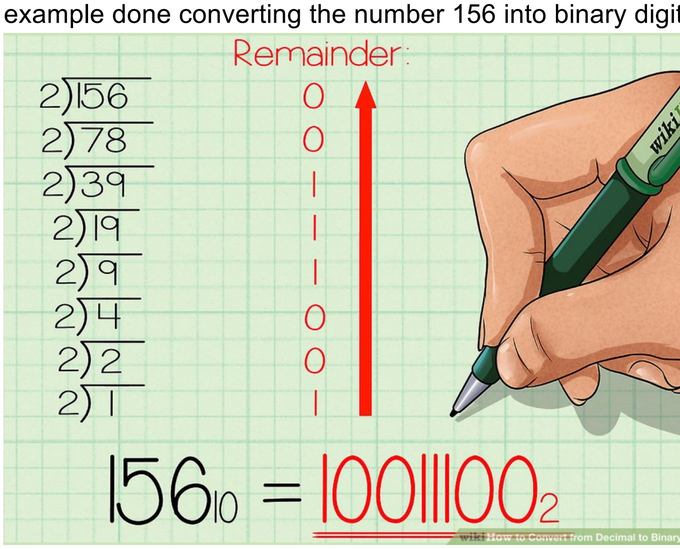
1. High cost.
2. DNA is significantly harder and slower to read than conventional computer transistors

Learning Goals

Through this practice, we hope to inspire students to associate synthetic biology with data science. we would like to give students an introductory understanding of basic data storage infrastructure, DNA string sequencing and encoding, and the workflow of storing data in DNA. By demonstrating the pros and cons of DNA data storage, we hope to illustrate the prospects of synthetic biology to students and motivate them to conduct research in the future.

Protocol

1. The students should prepare a string of texts that they wish to translate into binary sequences and DNA nucleotides strings.
2. Provide students with the ASCII table. For each character in the string of texts selected, they should find the corresponding Decimal number. Line the decimal numbers up in order. http://www.asciitable.com
3. For each decimal number, calculate the corresponding Binary sequence:
    a. Set up the number
    b. Divide the number by 2, create a column of remainders, and write the remainder (0 or 1) for this number down as the first element of this column. This should be set up preferably next to the dividend for better visualization.
    c. If the divided result is not 0 or 1, keep dividing it by 2, and write down the remainder (either a 0 or 1) downwards in the column, remember to keep the remainder to be in approximately the same line as the dividend.
    d. If the divided result is 0 or 1, simply continue downwards in the column, and write down your result as the end of this column.
    e. Once you got your column, start at the BOTTOM of the column and read upwards to the top. Now you have your binary sequence!
    f. For a more visual representation, you can take a look at this picture that demonstrated an example done converting the number 156 into binary digits:



4. For your binary sequence, divide them up by bits: two binary numbers at a time. Then convert all bits into DNA nucleotides according to the rule below.
    a. 00 -> A
    b. 01 -> G
    c. 10 -> C
    d. 11 -> T
5. Now you have your DNA nucleotide sequence encoded for your texts!