DNA Digital Data Storage

A Seminar Report

Submitted by

Deepak Raya Velgapuni (160115735091)

Bachelor of Engineering
From

Chaitanya Bharathi Institute of Technology
(Autonomous)

In

Electronics and Communication Engineering

Seminar Incharge:

Dr. A. Vani
Ms. Shakira Begum Shaik

ABSTRACT

The DNA is nature's hard drive; it has been used for storing instruction of cells, since the start of life on earth. Therefore, there is possibility of storing any kind of data into the genetic code.

Conventional mass-storage systems were doing the job cheaply and reliably. There was no compelling reason to seek out new options but, there is a problem, the data that we have generated in last 50 years is being generated in 50 minutes form past 2 years. To address this huge data deluge problem we have to search for compact and efficient storage systems and the help comes from DNA, which has a high density storage capability in low volumes.1g of DNA can store 216 petabytes of data. On the other hand, traditional mass-storage technologies are starting to approach their limits. With hard-disk drives, we're encountering a limit of 1 terabyte—1,000 GB—per square inch. Past that point, temperature fluctuations can induce the magnetically charged material of the disk to flip, corrupting the data it holds.

A glimpse into DNA data storage and various issues related to this such as, encoding, decoding, excerpts of research, cost, advantages and downsides are presented in this seminar.

CERTIFICATE

Certified that seminar work entitled "**DNA Digital Data Storage**" is a bonafide work carried out in the eighth semester by "**Deepak Raya Velgapuni**" in partial fulfillment for the award of **Bachelor of Engineering** in **Electronics and Communication Engineering** from **Chaitanya Bharathi Institute of Technology** during the academic year **2018-2019**.

SIGNATURE

NAME OF SEMINAR COORDINATOR

CONTENTS

CHAPTER 1

INTRODUCTION

DNA digital data storage is defined as the process of encoding and decoding binary data to and from synthesized DNA strands. DNA molecules are genetic blueprints for living cells and organisms. Although DNA data storage became a popular topic in the 21st century, it is not a modern-day idea. Its origins date back to 1964-65 when Mikhail Neiman, a Soviet physicist, published his works in the journal *Radiotehnika*. Neiman wrote about general considerations regarding the possibility of recording, storage, and retrieval of information on DNA molecules.

DNA can store remarkable amounts of genetic information (DNA is nature's hard drive).

The DNA molecule is a double-helix staircase of billions of molecular blocks, called base pairs, whose arrangement determines much of what makes each of us unique. About 3 billion base pairs are contained in a human body.

The basic idea of digital data storage in DNA is to covert the binary data into DNA nucleotides.

The motivation for this idea is to store the huge amounts of genomic information efficiently and with low power.

Conventional mass-storage systems were doing the job cheaply and reliably. There was no compelling reason to seek out new options but, the situation has changed drastically over the last 15 years. We face an unprecedented data deluge in medicine, physics, astronomy, biology, and other sciences. The Sloan Digital Sky Survey, for example, produces about 73,000 gigabytes of data annually. At the European Organization for Nuclear Research (CERN), the Large Hadron Collider

generates 50 million GB of data per year as it records the results of experiments involving, typically, 600 million particle collisions per second. These CERN results churn through a distributed computing grid comprising over 130,000 CPUs, 230 million GB of magnetic tape storage, and 300 million GB of online disk storage.

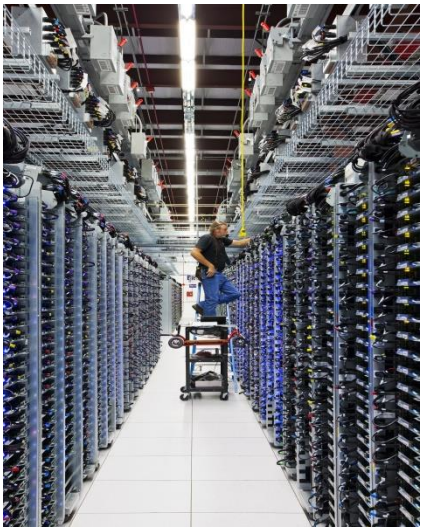
CHAPTER 2

WHY IN DNA?

The main advantage with DNA data storage is density and power.

1. Data density advantage:
    1. Lot of data can be stored into tiny bit of space.
    2. 1 g of DNA can approximately store 216 petabytes of data
    3. Which is not the case with conventional mass storage systems?
2. Shelf life advantage:
    1. Normal conditions half-life is around 500 years
    2. In dark and cold conditions it can be stored thousands of years
3. The power advantage:
    1. DNA storage doesn't require any power to maintain the data ,which would drastically reduce the costs electric power ,On the other hand the conventional data storage systems require a huge room with coolants and power systems providing electricity 24x7 .

    This makes the DNA the most efficient system, for growing amount of data.



CHAPTER 3

Big picture of DNA storage system

The Cell instructions inside DNA:

DNA made of four organic bases A, T, G and C, the DNA strand when cut into half, the sequence of bases determine the cell characteristics.

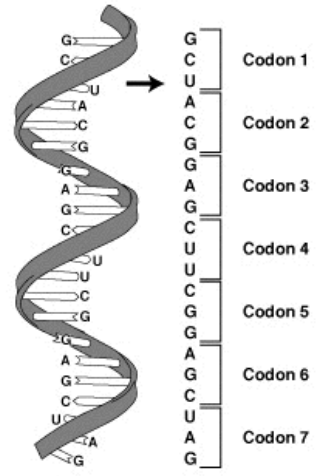
The specific sequence of these bases into groups of three called codons.

Codons construct the amino acids which gives instruction to make Proteins.

For the digital data, the codon kind of codes can be used to store other information too. First the binary data is converted to base 3 encoding and then to DNA alphabets using algorithms.

The DNA alphabets encoded message is then used to synthesize the DNA.

For retrieval of the stored information, the encoded DNA strand is surrounded with genetic markers. The above stand is sequenced using PCR and the message is decoded, if we have few redundant copies of information, we can correct mistakes automatically.



CHAPTER 4

Coding theory and DNA storage

DNA-based storage systems are new and uncharted territory for coding theorists, as the amount of data to store increase we need to augment with error control coding.

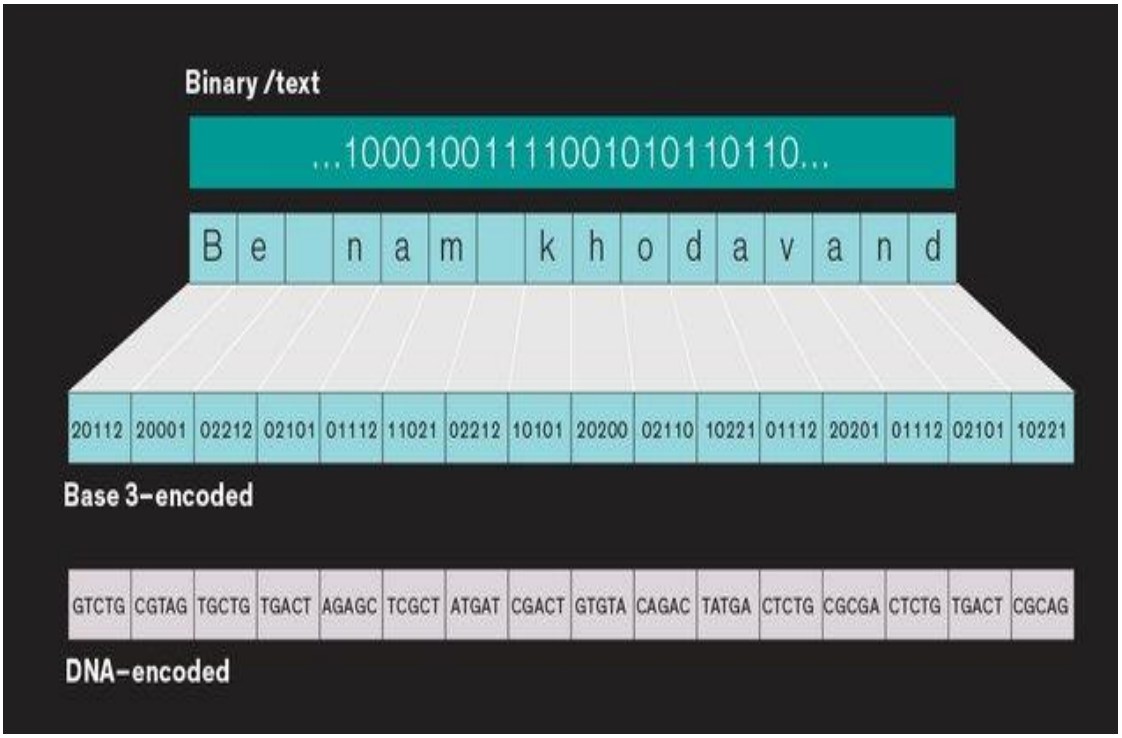
Types of encoding:

1. Base 3 encoding.
2. 2 bits per base encoding.

Base 3 encoding is found to more efficient with low error rate per alphabet.

Encoding data into a single long strand of DNA is asking for trouble when it comes time to recover the data. A safer process encodes the data in shorter strands. We then construct the first part of the next strand using the same data found at the end of the previous strand. This way we have multiple copies of the data for comparison.

Damerau distance codes, which in natural-language processing are used to catch errors like misspellings (for example, “smort” instead of “smart”), can identify the spots in binary code where 1s and 0s have likely been substituted by mistake during copying or transcription. Damerau distance codes can also be used to address the errors that occur in DNA, even though they’re more complex than binary errors. Sometimes bases are inadvertently deleted, and sometimes two will swap positions, errors that do not often occur in binary code.


Binary /text
...1000100111100101011010...
B e n a m k h o d a v a n d
20112 20001 02212 02101 01112 11021 02212 10101 20200 02110 10221 01112 20201 01112 02101 10221
Base 3-encoded
GTCTG CGTAG TGCTG TGACT AGAGC TCGCT ATGAT CGACT GTGTA CAGAC TATGA CTCTG CGCGA CTCTG TGACT CGCAG
DNA-encoded

CHAPTER 5

DNA sequencing and sequencing devices

DNA sequencing is the reading out of base pairs, which is retrieval of stored data. The cost for sequencing reduced a lot, and there are plenty of devices available in the market at present for sequencing. There are various methods for sequencing, the 2 most common methods are discussed below,

1. Shotgun sequencing
2. Nano pore-sequencing

"Shotgun-style" sequencing breaks copies of the long, unwieldy DNA strand into fragments of varying lengths. After those shorter segments are read, they can be compared with different fragments to reconstruct the entire sequence, although this method can introduce uncertainty about the placement of individual fragments.

Nanopore sequencers read long strings of DNA bases one by one, and because of the speed at which they do so, they will occasionally misread a particular base. Unlike the simple misreading of 1s and 0s, however, the odds of bases being mistaken for one another varies, due to their complex molecular structures and even the orientation the strand is in as it passes through the nanopore.

.

DNA SEQUENCER

CHAPTER 6

Disadvantages of DNA data storage

- Cost: It involves high cost to synthesize and sequence DNA .but it has drastically reduced compared to the past.
- Time: There is a delay in process of synthesis and sequencing DNA.
- Complexity: DNA isn't so simple. It's inherently unordered and lot of things to consider.

CHAPTER 7

Few excerpts of storing information into DNA

- 2012-scientist encoded 739 KB of computer files-154 Shakespeare sonnets and Martin Luther King Jr. "I have a dream" speech. (Ewan Birney and Nick Goldman(2012)).
- 2016-Microsoft research and University of Washington- stored 200MB of data

CHAPTER 8
CONCLUSION

- DNA based storage system can act as compact and efficient storage systems.
- Although the cost of DNA synthesizing and sequencing is high, it has reduced to one three-millionth what it was 10 years ago.
- There are already few excerpts of data storage in DNA with zero error in retrieval.
- There is a lot of research going behind DNA storage systems, probably in near future we might store the entire internet data (1.1 ZB) into few grams of DNA.

Making DNA-based storage a practical reality will require cooperation among researchers on the frontiers of synthetic biology and coding theory. We've made big strides toward realizing a DNA-based storage system, but we need to develop systems to efficiently access the information encoded into DNA. We need to design coding schemes that guard against both synthesis and sequencing errors. And we need to figure out how to do these things cheaply.

If we can solve these problems, nature's incredible storage medium—DNA—might also store our music, our literature, and our scientific advances. The very same medium that literally specifies who we are as individuals might also store our art, our culture, and our history as a species.

REFERENCES

- Olgica Milenkovic. (2018)- “Exabytes in a Test Tube: The Case for DNA Data Storage”-IEEE spectrum.
- Goldman, N. B.; Birney, E. (2013). "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA". *Nature*. 494 (7435): 77–80.