Feature extraction from disturbed algorithmic patterns for DNA data storage

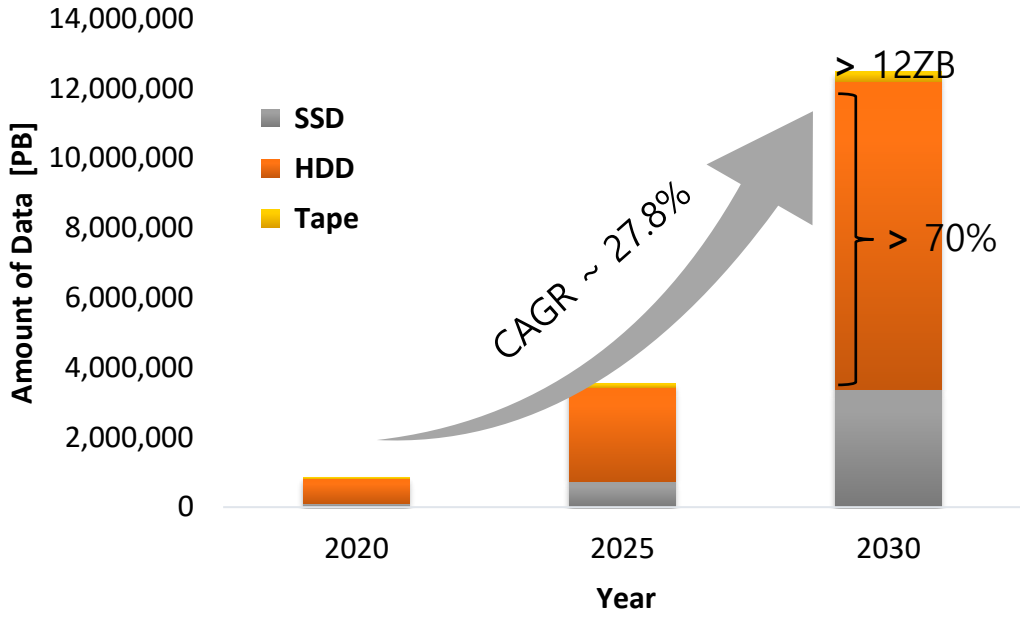
Presenter: Suyoun Park (Sungkyunkwan University)

Contents

- Introduction
- What is DNA data storage
- Retrieval in DNA data storage
- Conclusion

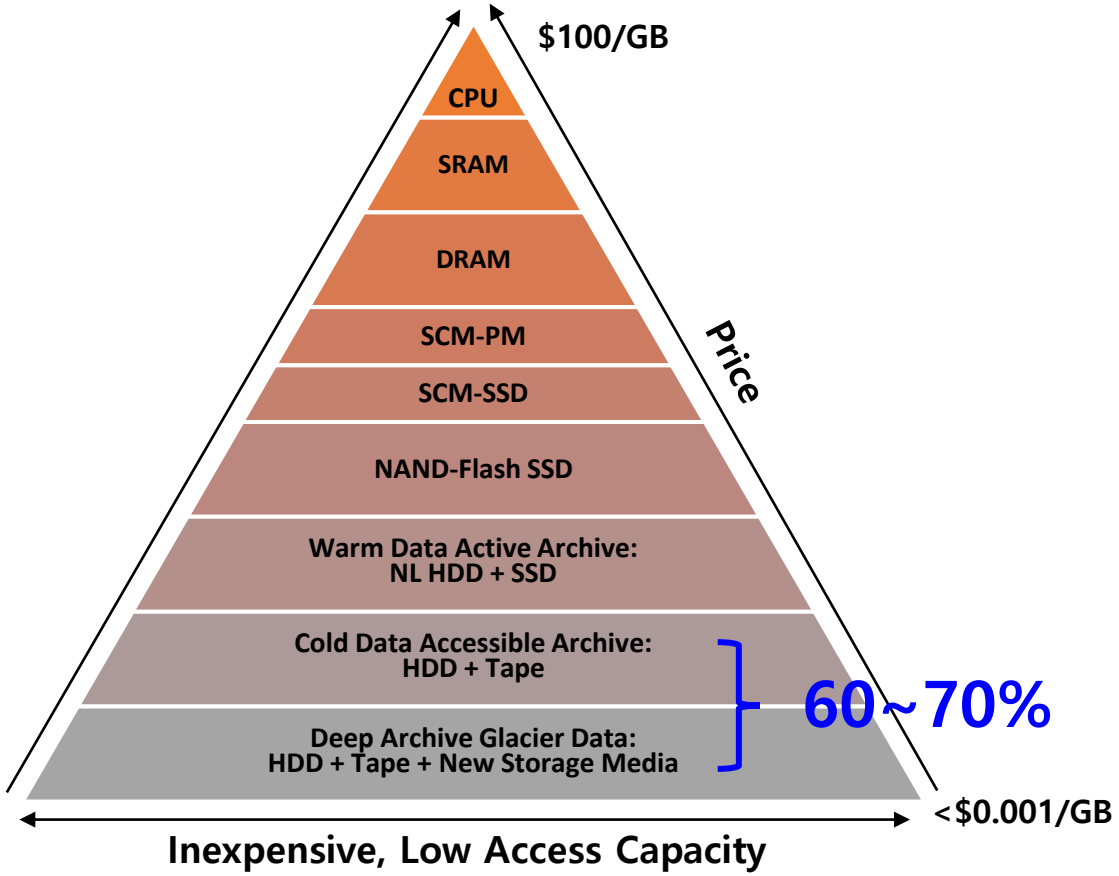
Introduction

- As Big data and AI evolve, countless data are created and stored around the world every day. As a result, the demand for data storage devices is increasing.
- Power, facility costs, and replacement costs due to storage life in the data center for storing data are challenges facing the near future.



Introduction

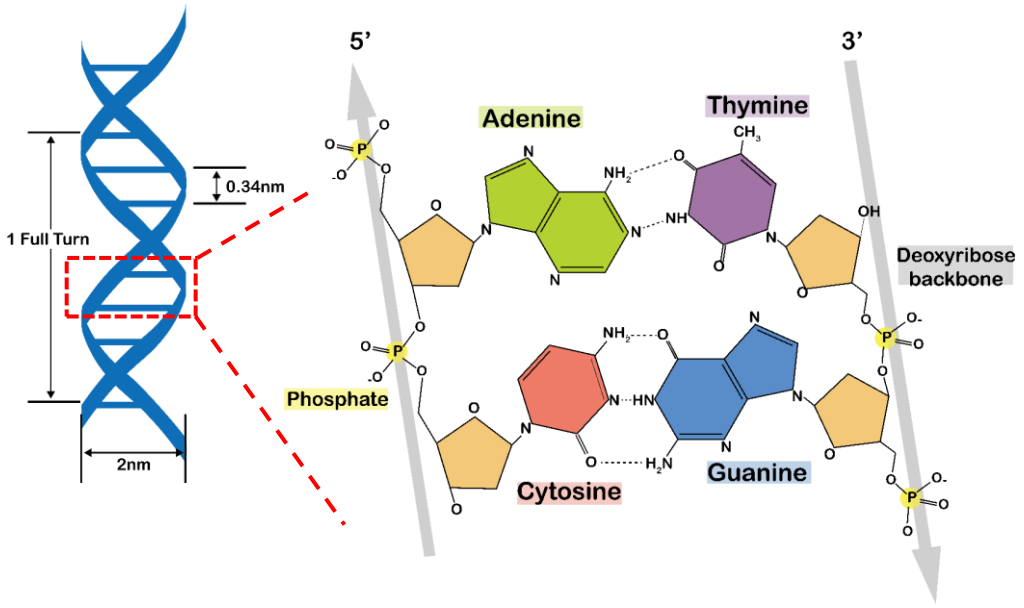
- HDD and Tape for Cold and Deep Archive with low frequency of data access account for 60-70%, and research to replace them with DNA Storage is being discussed.



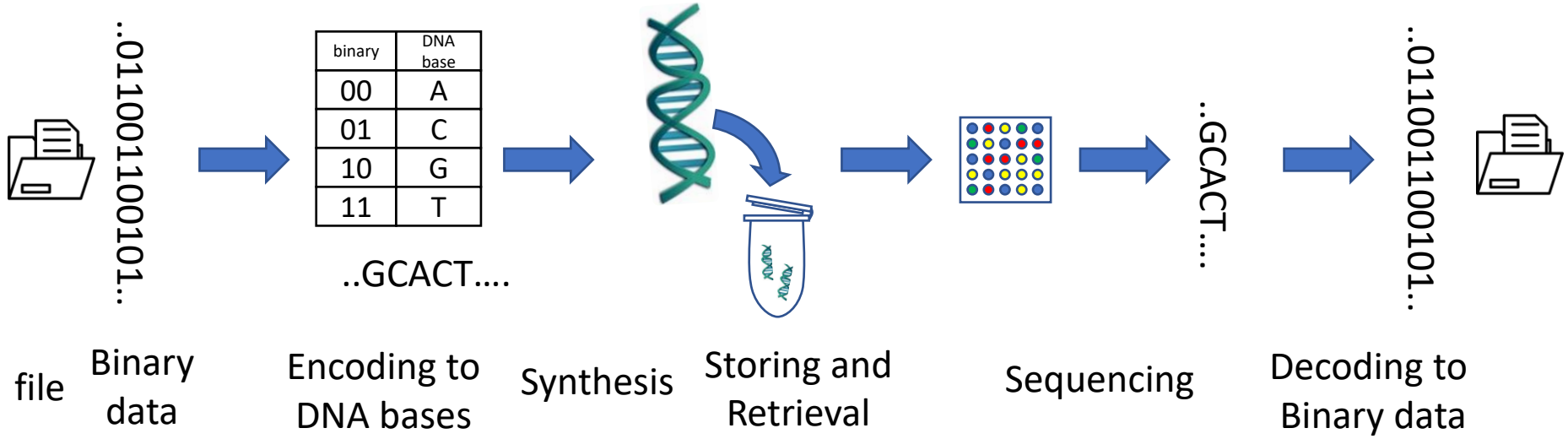
* ref : The Evolving Storage Pyramid@Gartner
- SRAM(Static random-access memory), DRAM(dynamic random-access memory)
- SCM(Storage class memory), SSD (Solid state drive), NL HDD(Nearline hard disk drive)
- Nearline = Near-Online

What is DNA data storage

- DNA
  - Deoxyribonucleic acid (DNA) is a molecule that carries genetic information used for the growth and development of all living things.
  - The two strands of DNA are called polynucleotides, which are made up of a complementary combination of four bases.
    - A(adenine), T(thymine), G(guanine), and C(cytosine).
  - Bases make pairing by hydrogen bonding, : A with T, C with G



What is DNA data storage



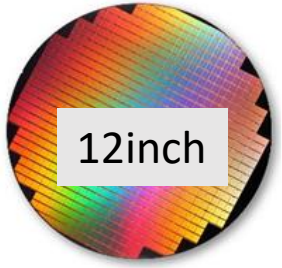
- DNA data storage is to store digital data in the base sequence of DNA : 00→A, 01→C, 10→G, and 11→T
- The synthesized DNA is stored and searched for the information want. A sequencing process is required to read the retrieved DNA, and DNA sequence is restored to a digital file after a decoding process.

What is DNA data storage

- Since a large capacity may be implemented in a small amount and the storage device may be stored at least 1,000yr, the replacement cost of the storage device may be significantly reduced.
- It is easy to replicate data using self-assembly characteristics, and there is little power consumption accordingly.

For storing 200PB of data

1TB HDD
200,000 (90t)



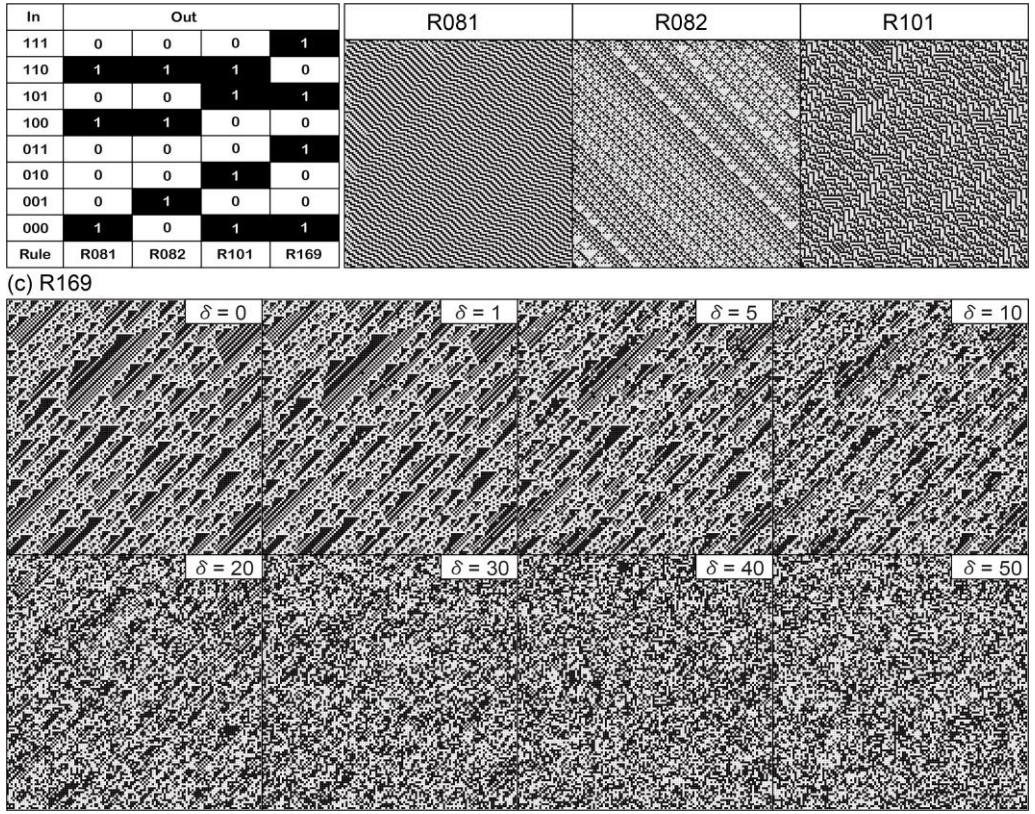
512GB of NAND
3,518 of wafers

DNA
1g of molecule

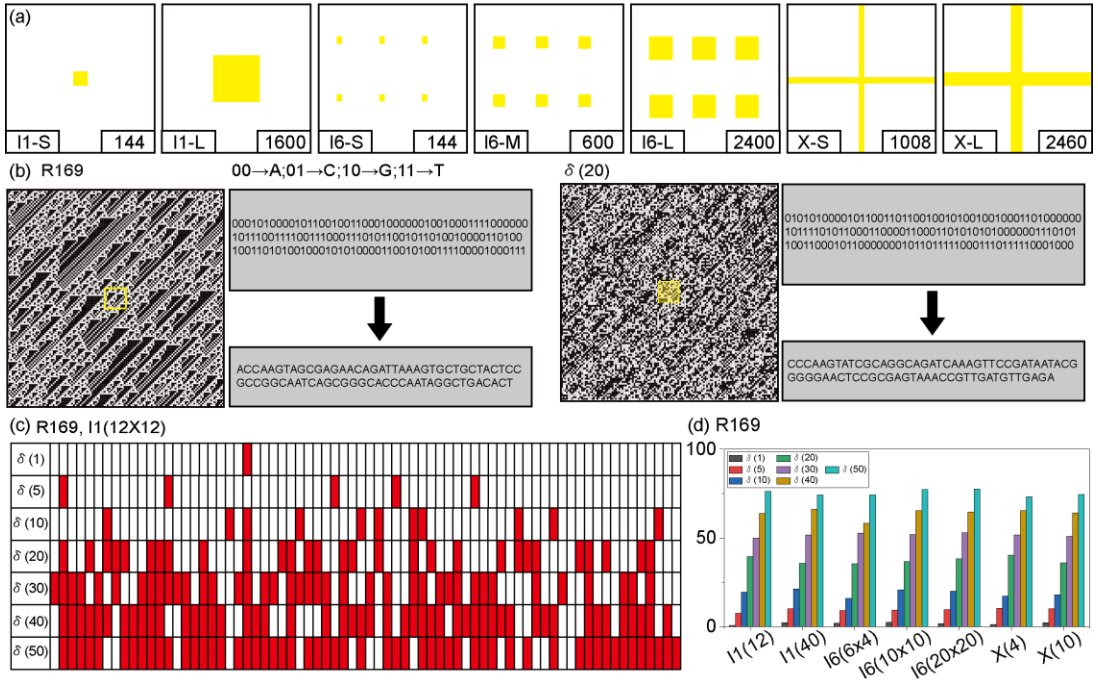
Retrieval in DNA data storage

- The pattern generation through 3-input/1-output logic implemented algorithmic pattern
- The disturbance ratio δ (n) indicates the number of cells that have added changing initial cells (i.e., 0 to 1 and 1 to 0) divided by the total number of cells

  (128 × 128 = 16,384 cells)
- The feasibility of retrieval experiment through extracting feature from algorithmic pattern with analysis of numerical parameters



Retrieval in DNA data storage



The feature representation from pattern and analysis of dissimilarity with δ = 0 and δ ≠ 0 of features

(a) The various type of features, which are represented to pattern as marked yellow shapes (one box, six boxes, and cross)

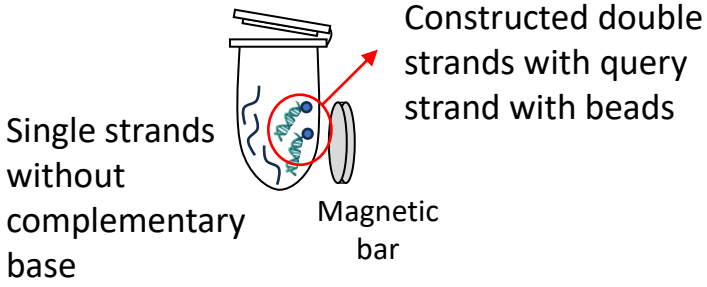
(b) The comparison to information bits and nucleotides placed in the box of δ (0) and δ (20)

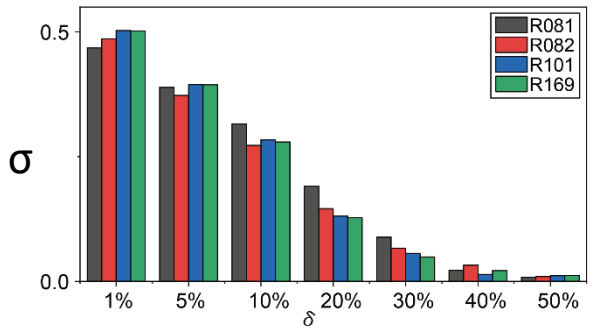
(c) The comparison for nucleotides placed at feature from pattern with disturbance

(d) The dissimilarity of feature types via disturbances

Retrieval in DNA data storage

- Experiments are conducted by using construction of double strand with query strand with beads.
- Strand with little dissimilarity can make double strand with query.
- Sequencing do extracted strand (little dissimilarity strand)
- A graph of standard deviation (i.e., $\sigma = [\sum(\alpha i - 0.5)2 / (N - 1)]1/2$, where αi is α of original and disturbance patterns at corresponding inputs, N is number of inputs (i.e., 111, 110, 101, 100, 011, 010, 001, and 000), respectively.) with respect to the various noise under the five different rules





Conclusion

- Using DNA data storage, it is possible to store huge amount of data in extremely small amount of DNA instead of using conventional data storage.
- Many researches are reported about DNA data storage in efficient encoding system from binary to DNA bases, Synthesis, Sequencing, storing and retrieval and error correction code, etc.
- In this presentation, retrieval experiment are applied numerical analysis by using feature in order to represent original picture.

Thank you

Suyoun Park