IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Novel Digital Data Storage Model: TFAM-DNA Packing Architectural Role for the Stability of DNA

Irene Choi[1], Kyunghwan Roh, Minkyung Sung, Soyoon Kim

1.Singapore American School, 40 Woodlands Street 41, Singapore, 738547, Singapore; gosyber1@gmail.com

ABSTRACT: DNA-based data storage is a rapidly evolving technology that could solve the rising energy and space needs of current information storage. However, DNA's stability limits its potential for future data storage systems. To address limitations to the current working storage method, we developed a novel method to store DNA in an aqueous solution with enhanced stability. Forming a Transcription Factor A Mitochondria (TFAM)-DNA complex in an aqueous solution, where TFAM encapsulates the DNA to protect it from UV irradiation and oxidation stress, could enhance the stability of DNA during storage and retrieval. First, we cloned the *TFAM* gene, overexpressed the TFAM protein, and purified the TFAM protein from *E. coli*. Next, we designed a smiley face pixel information in a black and white image (10x10). Then, we converted the data into binary numbers (white=0, black=1). The 10x10 binary numbers were then switched to nucleotides, with zero being either G or C and one being A or T. The image was encoded into nucleotides, synthesized into DNA, and cloned into a plasmid vector. After each stress was induced, the cloned DNA plasmid was sequenced and decoded back to the original image. Our result shows that the TFAM-DA complex provides a stable storage method even with harsh UV irradiation and oxidative stress conditions in aqueous solutions. In conclusion, our storage method's novelty allows for frequent access to stored data in an aqueous solution and provides protection from external stress.

KEYWORDS: Molecular Biology; Data Storage; TFAM; Binary Image; DNA protection.

Introduction

Digital data production has been growing exponentially, outpacing the growth of mainstream storage. Most data generated are discarded, but a portion still needs to be stored.[1] Thus, data storage centers have emerged to store massive amounts of data. Hard disk drives (HDDs) are considered a dominant storage technology due to factors such as the low price required for maintenance.[2] HDD stores and reads the binary code from a magnetic surface and rotating platters.[3] Another well-known method of data storage is solid state disks (SSDs). SDDs are currently argued to be a better storing method than HDDs due to low power usage.[4]

Data storage centers serve three primary purposes: computing, storing, and networking.[5] They process the data to operate the application, store the data with backups, and connect the data with different centers if necessary. Preserving the data long-term is critical for corporations requiring frequent recall of stored information. Notwithstanding, there are a few limitations of the current data storage model. First, data storage centers are not robust.[6] Most data storage centers operate on electricity; thus, the malfunctioning of electrical systems may lead to a significant loss of data. Moreover, data is currently stored at low information density, requiring larger space and more capital storage centers.[6]

Understanding the limitations of the current data storage methods, DNA has been proposed as the new alternative as a data storage medium. To store binary data in DNA, the data need to be converted into nucleotide sequences, each representing 0 and 1. DNA data storage models can be classified based on the duration of storage and the frequency of data recall.[7]

There are a few benefits of storing data in DNA. First, data storage in DNA can hold up to 1018 bytes per mm^3 -- six times denser than the densest storage method available. Moreover, unlike HDDs or SDDs, the DNA storage method enables swift data replication. [8]

Despite these various benefits, the DNA storage model still faces some limitations.[7] For example, stability has not been fully achieved. Data-digested DNA can be damaged by changes in storing conditions or exposures to other stress factors. UV light, for example, interrupts and alters the connections between nucleotide bases. Moreover, synthesized DNA cannot yet fully secure the data if the data exceeds the 10x10 binary form. Hence, its benefits are not utilized to their full potential, limiting the use of DNA as a primary data storage medium.

Histones are suggested as a solution for DNA stability, but there are a few limitations to histones. A histone octamer consists of smaller subunits, which bind together to protect the DNA.[9] This complexity of histones makes it a challenge to be used in DNA-based data storage. Furthermore, histones do not fully cover the DNA, failing to preserve data security.

Transcriptional Factor A Mitochondria (TFAM) was proposed as an alternative to histone. TFAM is a protein that covers mitochondrial DNA (mtDNA) to protect the DNA from external stress factors.[10] As TFAM is abundant, it can cover the entire region of mtDNA and display a histone-like behavior in the mitochondria.[11] Thus, we hypothesized that forming a TFAM-DNA complex in an aqueous solution, where TFAM encapsulates the DNA to preserve DNA from

DOI: 10.36838/v5i3.7

various stress factors, could enhance the stability of DNA during storage and retrieval.

To prove the hypothesis, purified TFAM was necessary. TFAM, collected from the PCR and the DNA vector inserted in the *E. coli*, was purified using Ni-NTA nanobeads. We then calculated the optimal ratio of TFAM to DNA by comparing different concentrations of TFAM solutions. The TFAM-DNA complex was formed using the optimal ratio and underwent multiple stress tests through exposure to hydrogen peroxide (H_2O_2) and UV light. The results validated the effectiveness of TFAM in securing DNA-based data.

Methods

PCR of TFAM mRNA:

TFAM mRNA coding sequence (CDS) was amplified via PCR. The following materials were used for the PCR reaction: distilled H_2O, forward primer, reverse primer, *TFAM* cDNA, the mixture of optimized DNA polymerase, dNTP, and buffer. The following PCR condition was used to amplify the target genes: 95 ˚C for 5 minutes, 30 cycles of 95 ˚C for 20 seconds, 58 ˚C for 30 seconds, and 72 ˚C for 45 seconds with a final extension at 72 ˚C for 5 minutes.

Agarose Gel Electrophoresis:

Agarose gel electrophoresis was then conducted to check if the correct protein could be produced with a size of 28 kDa. Agarose powder and DNA staining solution were mixed with Tris-acetate EDTA (TAE) to create the gel. DNA size markers were added on both ends of the gel. Samples collected through PCR were then added to different columns. Electric currents ran through to analyze the samples.

PCR Purification:

To remove residues other than TFAM, a PB buffer provided by a PCR purification kit (Enzynomics) was added to the test tube with the TFAM solution. After centrifugation, the first amount of excess liquid was removed, and the NW buffer was added for additional washing. Through centrifugation, other possible residues were again removed. The condensed TFAM was placed into a new 1.5 mL tube with an elution buffer for a final wash. The solution was centrifuged for one minute, and excess liquid was removed.

pET28 Vector Digestion and Ligation:

BamH1 and Xhol restriction enzymes were used to clone the TFAM gene into the pET28 vector. The DNA solution with enzyme cut buffer, BamH1 (Enzynomics), and Xhol (Enzynomics) was incubated for 10 minutes to provide sufficient time for enzymes to cut the amplified TFAM gene. A PCR purification was then performed to eliminate the BamH1 and Xho1 restriction enzymes. The individually cut protein and vector were then connected through ligation. First, a ligation buffer was mixed with a digested vector, digested DNA insert, and DNA ligase. After being vortexed, the new solution was centrifuged to remove the buffer. The remaining sample was incubated at 15 ˚C for 4 hours to provide time for the vector and the protein to connect.

pET28 Vector Transformation:

The pET28-TFAM vector was transformed into BL21 (DE3) *E. coli* strain with a HIS-TAG-producing sequence in the N-terminus. First, the ligated DNA mixed with the *E. coli* was heat-shocked for 90 seconds at 42˚C, facilitating the entry of DNA into *E. coli*. Then, to select the *E. coli* containing vectors with TFAM, kanamycin was added to the *E. coli* solution.

Cell harvesting and sonication:

The BL21 (DE3) cell harvest allows the target protein to be collected. First, to test whether IPTG serves its purpose in removing the lac repressor, two tubes were prepared for each sample, one with IPTG and one without IPTG. Once the sample was fully swirled, the sample was collected. After centrifugation, PBS was added to the samples to loosen the cells that were clumped together. SDS loading dye was added for sonification, which ran for 1 minute for each sample. The power was set to 1%, and the pulse was on for 1 second and off for 10 seconds to prevent protein denaturation. Then, the samples were centrifuged for 1 minute to remove any possible leftover solution from the tube.

TFAM Protein Purification using Ni-NTA magnetic nanobeads:

Pre-included bacterial culture with IPTG was prepared and centrifuged, allowing the supernatant to be discarded. After centrifugation, the harvested cell was resuspended in a washing buffer. The resuspended cells were then sonicated twice. Then, *E. coli* lysate was centrifuged for 5 minutes to remove the supernatant from the solution. The next part of purification was equilibrating Ni-NTA magnetic nanobeads. 1 mL of Ni-NTA was transferred to a tube placed under the Neodymium (Nd) magnet for 1 minute. After removing the supernatant, the sample was equilibrated by mixing 1 mL of binding buffer with the bead slurry. After the tube was placed on the Nd magnet for another minute, the remaining liquid was removed. This step was repeated four times in total. Lastly, to separate TFAM from Ni-NTA nanobeads, the sample was incubated for 1 minute at room temperature with the elution buffer and was placed on the Nd magnet for 1 minute. The elution sample was then removed as a supernatant. The process was then repeated to ensure the purification of TFAM. The results were collected through SDS-PAGE gel.

TFAM Protein Quantification:

Coomassie Brilliant Blue G-250 (CBBG), a dye, was used to stain the protein. Once the reaction finished, the protein changed from maroon to blue. In a 1.5 mL tube, the BSA and distilled water were mixed to create a BSA standard curve sample and calculate the density of the solution. 195 μL of the diluted Bradford reagent and 5 μL of BSA standard sample curve solution was added, and the unknown sample in to match a total volume of 200 μL. The sample was mixed with the reagent and was left for 5 minutes. Then, the microplate reader was used to record the absorbance at a wavelength of 595 nm. The data was then plugged into Beer's Law to quantify the protein in the solution.

TFAM-DNA Binding Test:

The TFAM-DNA binding test serves the purpose of determining the optimal ratio of TFAM: DNA vector. Nine tubes were prepared, and 3 μL of DNA was put into each test tube. Starting from 0 μL, 1 μL of TFAM increased per test tube. The total volume of the solution was then matched to 11 μL

by adding the elution buffer. DNA loading dye (2 μL) was added at the end to load the gel.

Stress Test:

Two stress tests were performed to assess how successful TFAM was in protecting the data-stored DNA. The first experiment was a 3 M H_2O_2 test. Two different tubes were prepared, each labeled as 'Elution, H_2O_2' and 'TFAM, H_2O_2'. 3 μL of DNA, 8 μL of the elution buffer, 1 μL of 3 M H_2O_2, and 2 μL of DNA loading dye were added to the tube 'Elution, 3M H_2O_2.' Everything except for the elution buffer, replaced by TFAM, was added to the tube 'TFAM, 3M H_2O_2.' The second stress test done on the TFAM-DNA complex was a UV test. Two tubes were prepared, each labeled as 'Elution, UV' and 'TFAM, UV.' 3 μL of DNA, 8 μL of the elution buffer, and 2 μL of DNA loading dye were added to tube 'Elution, UV,' and 3 μL of DNA, 8 μL of TFAM, and 2 μL of DNA loading dye were added to the sample 'TFAM, UV.' Both samples were left for 5 hours.

Results and Discussion

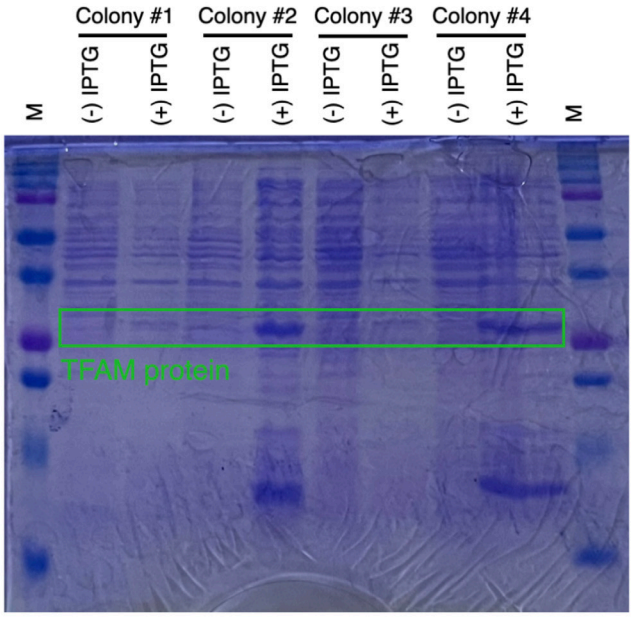

Figure 1: IPTG successfully induces TFAM protein in *E. coli*. The green box shows the location of TFAM protein, and it shows that colonies 2 and 4 with IPTG solution produced the most TFAM proteins. The SDS-PAGE was conducted after collecting the proteins from the harvested *E. coli* solution.

After cloning the TFAM gene, we transformed the plasmid vector to *E. coli* to overexpress the TFAM protein. Since IPTG induces the TFAM expression in the pET28 vector, the IPTG induction test was conducted to check if IPTG can induce the expression of the TFAM protein. After preparing four *E. coli* colonies containing TFAM plasmid, SDS-PAGE gel analysis was performed. The result indicated that colony # 2 and colony # 4 *E. coli* with IPTG produced the TFAM protein, whereas those without IPTG did not (Figure 1). This led to the conclusion that IPTG successfully induced the TFAM expression in *E. coli*.

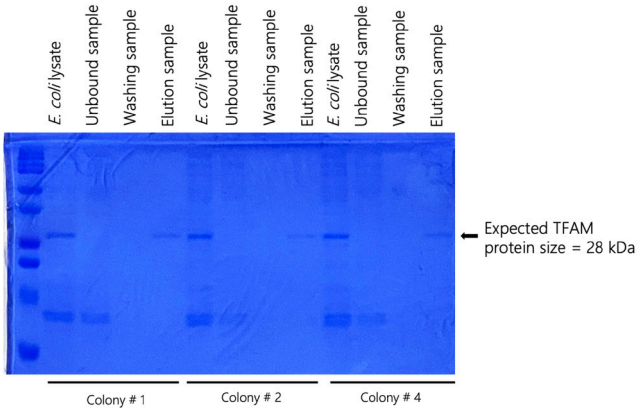

Figure 2: TFAM protein was successfully purified using Ni-NTA magnetic bead. Four samples from three colonies were analyzed: *E. coli* lysate, unbound sample, washing sample, and elution sample. SDS-PAGE gel visualizes the protein bands in each sample.

The *E. coli* lysate that contained the TFAM protein was collected to be purified. The lysate was purified using the property that a nickel column, his-tag at the end of TFAM protein, is attracted to Ni-NTA magnetic nanobeads. The protein was purified by binding the protein to the magnet, washing magnetic nanobeads, and eluting target proteins. One single band of TFAM (28kDa) was indicated in the SDS-PAGE gel, indicating the effectiveness of the purification (Figure 2).

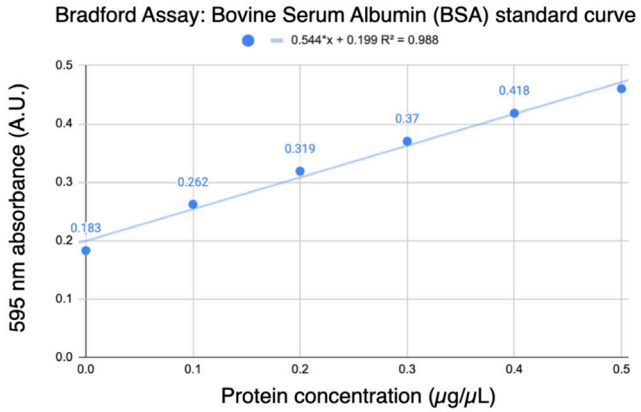

Figure 3: The standard curve of the Bradford assay using Bovine Serum Albumin (BSA) shows a high correlation between absorbance and protein concentration. The linear blue graph shows the standard curve of BSA protein. The microplate reader recorded the absorbance of BSA protein at a wavelength of 595 nm. The equation of the fitted line is indicated above the standard curve.

After checking the degree of purification, the amount of TFAM protein produced (concentration) was identified via Bradford Assay. The absorbance of Coomassie Brilliant Blue G-250 (CBBG) to TFAM was quantified using a micro-spectrometer. By comparing with the BSA protein, TFAM protein concentration was calculated. As a result, TFAM protein concentration was 0.235 μg/μL and 8.38 μM (Figure 3). Then, through the TFAM-DNA binding test, it was also concluded that the best TFAM:pSmile (2087 bp) binding mol ratio is 113.47:1.

DOI: 10.36838/v5i3.7

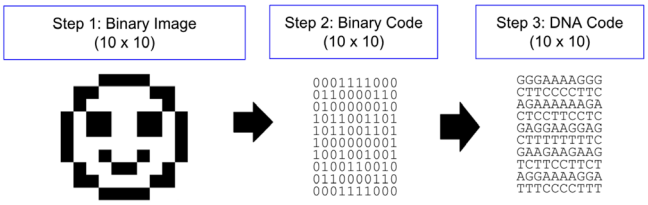

Figure 4: Converting binary image (smiley face) to the DNA sequence. 0 = white, 1 = black in Binary code. G or C = 0, A or T = 1 in DNA sequence.

A binary image (smiley face) is a digital image composed of 2 colors (black and white). Therefore, it is possible to represent a binary image as a binary code representing 0 = white and 1 = black. In this research, the smile face image was converted into binary code. Then, binary code was again converted into DNA code, where G or C represents binary code 0, and A or T represents binary code 1. Then, DNA was synthesized and cloned into the pBHA vector (pBHA/smile).

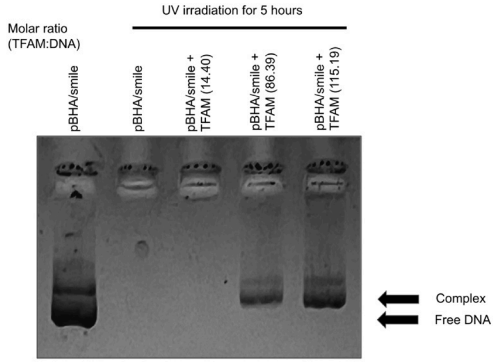

Figure 5: TFAM protects the DNA against UV irradiation. Agarose gel electrophoresis of pBHA/smile (lane 1-2), and pBHA/smile + TFAM (lanes 3-5). Lanes 2-5 were treated with UV irradiation for five hours.

Naked pBHA/smile (lane 1) was entirely disintegrated by an aggressive UV irradiation exposure (lane 2, Figure 4). After forming the TFAM-DNA complex with the molar ratios 86.39 and 115.19 (lane 4-5), the partial DNA remained even after the UV stress for five hours. However, a TFAM-DNA complex with a low molar ratio of TFAM (14.40) was insufficient to protect DNA from UV irradiation (lane 3). Moreover, the band intensity of DNA (lanes 4-5) was relatively low compared to the intensity of naked pBHA/smile (lane 1) without stress. This result suggests that the TFAM-DNA complex protects the DNA from UV irradiation, but it is insufficient to protect 100% of the DNA molecules in the sample.

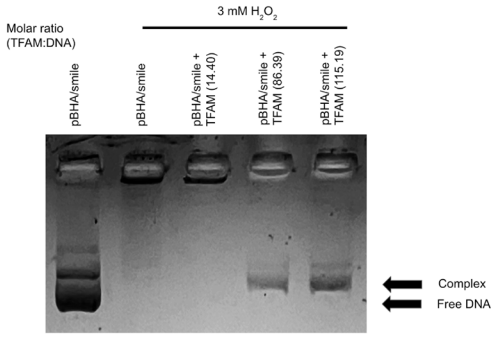

Figure 6: TFAM protects the DNA against H_2O_2 stress. Agarose gel electrophoresis of pBHA/smile (lane 1-2), and pBHA/smile + TFAM (lane 3-5) were treated as described in the protocol. Lane 1 to 5 from left to right.

Naked pBHA/smile (lane 1) is entirely disintegrated by aggressive H_2O_2 stress (lane 2, Figure 5. After forming the TFAM-DNA complex with a molar ratio of 86.39 and 115.19 (lane 4-5), the plasmid DNA remained even after the 3 mM H_2O_2 stress for five hours. However, a TFAM-DNA complex with a low molar ratio of TFAM (14.40) was insufficient to protect DNA from H_2O_2 stress (lane 3). As the UV stress data presented, the band intensity of either the molar ratio of 86.39 or 115.19 (lane 4-5) was relatively low compared to the band intensity of naked pBHA/smile (lane 1). Therefore, like the UV irradiation test, the result indicated that TFAM protects DNA from H_2O_2 stress but does not protect 100% of DNA molecules in the sample.

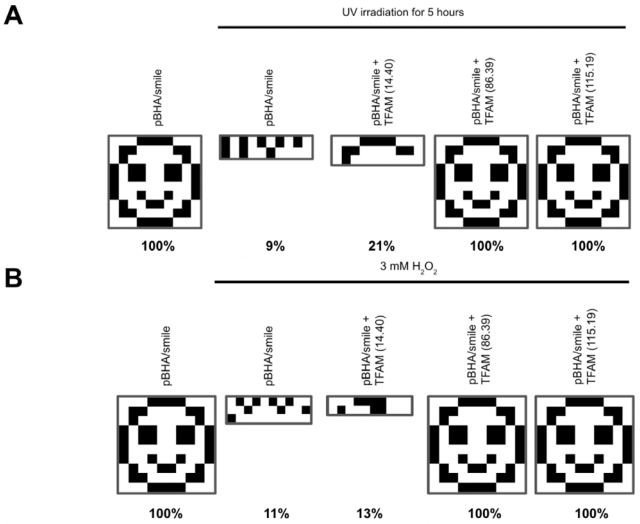

Figure 7: After DNA stress was induced, the DNA was sequenced and decoded to retrieve the original image. (A) UV irradiated DNA (B) H_2O_2 stressed DNA. The intactness of the DNA was measured by how much of the original image had been preserved after each stress had been induced. The percentage indicates the average recovery rate of the original image (n= 3).

Finally, to test whether TFAM protein can effectively protect data-stored DNA from various damages like UV light and H_2O_2, each TFAM-DNA complex was exposed to UV irradiation and H_2O_2 for 5 hours. After applying UV and H_2O_2 stress, the DNA sequence was analyzed by Sanger sequencing to check the DNA integrity in each sample. Then, the sequence was converted into binary code and decoded back to the original image.

In both UV irradiation and H_2O_2 stress, pBHA/smile presented low recovery rates: 9% and 11%, respectively (Figure 6). pBHA/smile with 1 μg/mL TFAM also displayed a low recovery rate in both UV irradiated and H_2O_2 stressed samples: 21 % and 13%, respectively. However, pBHA/smile with TFAM (Molar ratio of 86.39 and 115.19) recovered 100% of the original image even under stress conditions. This result indicates that the TFAM-DNA complex protects the DNA even in aggressive aqueous storage conditions. The results also reported that non-biological information, the smiley face image stored in plasmid DNA, can be retrieved without errors.

Conclusion

Among the many forms of molecular data storage, DNA is attractive. Data storage in DNA can store up to 1018 bytes per

mm^3, which is about six orders of magnitude denser than the densest storage method available today. Another unique advantage to DNA as a storage media is the ease of replicating DNA using PCR. This allows scientists to copy large amounts of data in a short amount of time and at a low cost. Lastly, DNA and DNA sequencing have a bright prospect, considering their increasingly crucial role in biological and medical research. The cost and time of sequencing will likely be improved in the future.

Even though data storage in DNA has many advantages over modern electromagnetic data storage, information stability in aqueous storage remains an important issue. Although the DNA is more stable than other encoded biopolymers, such as RNA, unprotected DNA is susceptible to UV irradiation and hydrolysis. DNA repair system reverses the loss of information in living organisms. Without these repair systems, ROS and ionizing radiation constantly degrade DNA. However, if DNA can be isolated and protected from stress, it can maintain its information over thousands of years, as ancient fossils stored DNA in plant seeds or bacterial spores. In this research, we designed a TFAM-DNA complex to store non-biological information (smiley face image) in DNA. It mimics the histone-like structure present inside the nucleus. It protects DNA from aggressive UV irradiation and high concentration of H_2O_2 conditions. Also, it can be both molecularly synthesized and disintegrated in water. These characteristics make TFAM-DNA complex an ideal assembly material for an aqueous storage system to protect DNA and retrieve information.

Acknowledgments

We want to thank Prof. Lee for his guidance in this research.

References

1. Li, L.; Novillo-Ortiz, D.; Azzopardi-Muscat, N.; Kostkova, P. Digital Data Sources and Their Impact on People's Health: A Systematic Review of Systematic Reviews. *Front. Public Health* **2021**,9, 645260.
2. Zhang, Y.; Liu, J. Prediction of Overall Energy Consumption of Data Centers in Different Locations. *Sensors* **2022**, *22* (10).
3. Passarino, G.; De Rango, F.; Montesanto, A. Human Longevity: Genetics or Lifestyle? It Takes Two to Tango. *Immun. Ageing* 2016 , *13* (1), 12. https://doi.org/10.1186/s12979-016-0066-z.
4. Lee, S.; Min, H.; Yoon, S. Will Solid-State Drives Accelerate Your Bioinformatics? In-Depth Profiling, Performance Analysis, and Beyond. *Brief. Bioinformatics* **2016**, 17 (4), 713–727.
5. Waithira, N.; Mutinda, B.; Cheah, P. Y. Data Management and Sharing Policy: The First Step towards Promoting Data Sharing. *BMC Med*. **2019**, 17 (1), 80.
6. Wang, W.; Krishnan, E. Big Data and Clinicians: A Review on the State of the Science. *JMIR Med. Inform*. **2014**, 2 (1), e1.
7. Dong, Y.; Sun, F.; Ping, Z.; Ouyang, Q.; Qian, L. DNA Storage: Research Landscape and Future Prospects. *Natl Sci Rev* **2020**, 7 (6), 1092–1107.
8. Ceze, L.; Nivala, J.; Strauss, K. Molecular Digital Data Storage Using DNA. *Nat. Rev. Genet*. **2019**, 20 (8), 456–466.
9. Young, N. L.; Dimaggio, P. A.; Garcia, B. A. The Significance, Development and Progress of High-Throughput Combinatorial Histone Code Analysis. *Cell. Mol. Life Sci*. **2010**, 67 (23), 3983–4000.
10. Chew, K.; Zhao, L. Interactions of Mitochondrial Transcription Factor A with DNA Damage: Mechanistic Insights and Functional Implications. *Genes (Basel)* **2021**, 12 (8).
11. Ngo, H. B.; Kaiser, J. T.; Chan, D. C. The Mitochondrial Transcription and Packaging Factor Tfam Imposes a U-Turn on Mitochondrial DNA. *Nat. Struct. Mol. Biol*. **2011**, 18 (11), 1290– 1296.

Authors

Irene Choi is a junior at Singapore American School interested in pharmacy and bioengineering. She is leading a biology lab club at school and has gained related experiences through iGEM, an international synthetic biology competition.

Minkyung Sung is a senior at the American School of Johannesburg, South Africa. She is interested in biochemistry and bioengineering. She has participated in research programs and clubs related to this field and gained more practical experience through iGEM, an international synthetic biology competition.

Kyunghwan Roh is a sophomore at St. Johnsbury Academy Jeju who wants to major in biotechnology and is interested in genetic engineering technology. He is the head of the science news club at school and has gained much experience in iGEM, an international synthetic biology competition.

Soyoon Kim is a sophomore at Newton South High School in Massachusetts, United States, who is interested in children psychology and wants to major in psychiatry. She has been involved in psychology and neurology programs and conferences and gained more experience though iGEM, an international synthetic biology competition.

DOI: 10.36838/v5i3.7