



DNA 데이터 저장 분야 최신 동향 보고서



임성순

KAIST 공학생물학대학원 교수

1

배경과 필요성

요즘 데이터는 정말 빠른 속도로 늘고 있습니다. AI 서비스가 늘고, 사물인터넷(IoT) 센서가 곳곳에 붙고, 초고해상도 영상과 디지털 트윈 같은 기술이 확산되면서 사람과 기계가 만들어내는 데이터 양 자체가 구조적으로 커지는 흐름입니다. IDC는 전 세계 데이터 총량이 2025년에 약 175ZB(제타바이트) 규모가 될 것으로 전망했는데, 이런 추세가 계속되면 데이터센터는 단순히 저장 장치를 더 사는 수준을 넘어 전력, 냉각, 공간, 유지보수 비용까지 함께 크게 늘어날 수밖에 없습니다. SNIA의 2025년 기술 리뷰는 이 상황을 “저장할지 버릴지”의 딜레마로 설명합니다. 저장할 데이터는 계속 늘어나는데, 지금의 HDD/SSD 같은 저장 장치는 수명이 짧아서(몇 년 단위) 주기적으로 장비를 교체하고, 데이터를 새 장비로 옮기는 작업을 반복해야 합니다. 이 과정이 장기적으로 비용과 운영 부담을 키운다는 것입니다.

이런 맥락에서 DNA 저장은 “당장 SSD/HDD를 대체하는 기술”이라기보다, 아주 오래 보관해야 하지만 자주 꺼내 보지 않는 데이터(아카이브/콜드 데이터)를 대상으로 먼저 적용 가능성이 큰 기술로 정리됩니다. 장기 저장 기술에서 중요한 조건은 대체로 네 가지입니다. (1) 아주 높은 저장 밀도(작은 부피에 많은 정보), (2) 상온에서 오랜 기간 안정성, (3) 저장 중에는 전력을 거의 쓰지 않는 ‘보관 중 무전력’ 특성, (4) 장비를 자주 갈아끼우지 않아도 되는 장기 운용 구조입니다. DNA는 분자 수준에서 정보를 담기 때문에 이론적으로 매우 높은 밀도를 가질 수 있고, 건조·차광·격리 같은 조건을 갖추면 장기 보관이 가능하다는 장점이 있습니다. 또 DNA를 읽고 쓰는 기술은 생명과학과 의료 산업이 커질수록 계속 발전할 가능성이 높아, “앞으로도 읽힐 포맷”이라는 점도 강점으로 꼽힙니다.

2

코덱과 시스템 설계

DNA 저장을 이해할 때 가장 중요한 점은, 이게 단순히 “DNA를 합성해서 넣어두는 기술”이 아니라는 것입니다. 실제로는 디지털 데이터를 DNA로 바꾸는 규칙(코덱, codec)과, 다시 원래 데이터로 되돌리는 규칙이 시스템 성능을 크게 좌우합니다. 코

덱은 쉽게 말해 “0과 1로 된 파일을 A/C/G/T 염기서열로 바꾸고, 다시 복원하는 방법”입니다.

코덱이 중요한 이유는 DNA 저장 과정에서 오류가 생기는 방식이 일반 저장장치와 다르기 때문입니다. DNA를 합성하고, 증폭(PCR)하고, 보관하고, 시퀀싱으로 읽어오는 과정에서는 단순한 ‘글자 하나 바뀔’뿐 아니라, 글자가 빠지거나(결실), 추가되거나(삽입), 특정 조각이 통째로 사라지는(드롭아웃) 문제가 생깁니다. 또한 PCR 때문에 어떤 DNA 조각은 과도하게 늘어나고 어떤 조각은 상대적으로 줄어드는 증폭 편향도 생길 수 있습니다. 그래서 코덱은 이런 “DNA 저장에서 흔한 오류 패턴”을 전제로 설계되어야 합니다.

이 때문에 데이터 표현 방식도 다양하게 발전했습니다. 예를 들어 가장 직관적인 이진 표현은 염기 하나에 2비트 정보를 담을 수 있지만, 같은 염기가 길게 반복되는 서열(호모폴리머)이 생기기 쉬워 실험적으로 불리할 수 있습니다. 반대로 삼진(ternary) 방식은 반복을 피하도록 설계되어 안정성이 좋아질 수 있지만, 염기당 정보 밀도는 낮아집니다(대략 1.58비트 수준). 이 외에도 “전이(transition)로 정보를 담는 방식”, “짧은 조각을 조합해 길게 만드는 방식”, “DNA에 니킹(절단) 같은 구조 변형으로 기록하는 방식”, “나노포어에서 읽기 쉬운 구조를 의도적으로 만드는 방식” 등 다양한 변형이 제안되어 왔습니다. 한마디로, 합성과 읽기 기술이 발전할수록, 그에 맞춰 데이터 표현(코덱)도 같이 진화하고 있는 것입니다.

오류 정정(ECC, error correction)은 이 코덱 위에서 작동합니다. 대표적으로 DNA Fountain 같은 방식은 일부 조각이 사라져도 전체 파일을 복원할 수 있도록 설계되어 “드롭아웃”에 강점을 보였습니다. 최근에는 실험 데이터를 기반으로 DNA 저장 채널을 더 정량적으로 이해하려는 연구가 늘었고, 이를 바탕으로 “어떤 오류가 어느 단계에서 얼마나 생기는지”를 반영해 코덱을 더 현실적으로 설계하는 방향으로 가고 있습니다. 또한 합성·시퀀싱 노이즈를 모사하는 시뮬레이터를 이용해 미리 설계를 검증하는 방식도 늘고 있습니다. 상용 관점에서는 결국 코덱이 단독으로 존재하는 게 아니라, 합성·회수·읽기 과정까지 포함한 운영 프로토콜과 함께 묶여 발전하는 흐름이 뚜렷합니다.

3

쓰기(합성) 기술 동향

DNA 저장에서 가장 큰 현실 장벽은 여전히 “쓰기(합성)가 비싸고 느리다”는 점입니다. 초기 시연 연구들은 DNA 저장이 가능하다는 것을 보여줬지만, 비용과 처리량 측면에서는 기존 테이프나 디스크 기반 아카이브와 비교가 어려운 수준이었습니다. 그래서 이후 연구의 큰 방향은 “어떻게 하면 대량으로 더 싸게 쓸 수 있는가”에 맞춰져 왔습니다. 예를 들어 Antkowiak 등은 광유도 기반의 대규모 병렬 합성(칩/어레이 방식)과 고도화된 복원 알고리즘을 결합해, 오류율이 높은 환경에서도 데이터를 다시 복구할 수 있는 방법을 보여주며 저비용 쓰기 방향을 구체화했습니다.

한편 최근 주목받는 흐름은 효소 기반 DNA 합성(EDS, enzymatic DNA synthesis)입니다. 기존 화학 합성은 성숙한 기술이지만 공정·폐기물·비용 측면에서 한계가 있습니다. 효소 합성은 수용액 환경에서 진행될 수 있어 지속가능성 측면에서 매력이 있고, 장기적으로는 더 빠르고 확장 가능한 합성 방식이 될 가능성이 있습니다. 예를 들어 Lee 등은 TdT 효소를 이용해 “템플릿 없이” DNA를 합성하는 접근을 보였고, Palluk 등은 효소-뉴클레오타이드 결합체를 이용해 염기를 한 단계씩 제어하려는 방향을 제시했습니다.

다만 효소 합성이 실제 상용화로 이어지려면 넘어야 할 산이 있습니다. 핵심은 정확한 염기 단위 제어, 삽입/결실 같은 오류 패턴의 관리, 어레이화(대규모 병렬 합성), 그리고 품질을 객관적으로 평가할 표준 지표입니다. 즉 “길게 만들 수 있느냐”만이 아니라, 얼마나 많이(throughput), 얼마나 빨리(latency), 얼마나 안정적으로(재현성), 자동화 공정으로 구현 가능하냐까지 함께 봐야 “상용 준비도”를 논할 수 있습니다.

4

보관과 랜덤 액세스

DNA를 장기간 보관하려면 DNA 자체의 안정성만큼이나 ‘어떻게 포장해서 보관하느냐’가 중요합니다. 예를 들어 실리카 캡슐화는 DNA를 산소·수분·자외선으로부터 격리해 안정성을 높이는 대표적인 방법입니다. 장기 저장 매체로 인정받으려면 이런 보관

방식이 실제로 얼마나 오래 버티는지(내구성)도 표준화된 방식으로 평가될 필요가 있습니다. 또 하나의 큰 관문은 랜덤 액세스(random access)입니다. DNA 저장에는 보통 수많은 DNA 조각이 한 용액(pool)에 섞여 있는 형태가 됩니다. 여기서 특정 파일만 꺼내 읽으려면, 단순히 “전부 시퀀싱해서 찾는 방식”은 대용량에서 비현실적입니다. 시퀀싱에는 한 번에 읽을 수 있는 양(read budget)이 있기 때문입니다. 그래서 파일이나 블록마다 주소(address) 역할을 하는 프라이머(primer)를 붙여두고, PCR로 필요한 조각만 선택적으로 증폭해 꺼내는 방식이 널리 쓰입니다. 다만 PCR 기반 랜덤 액세스는 단점도 있습니다. SNIA가 언급하듯 보통 파괴적(한 번 쓰면 원본이 줄어드는 형태)일 수 있고, 회수 과정에서 1-2시간 수준의 추가 지연시간이 생기기도 합니다. 최근 Boolean search 접근은 단순히 “파일 하나 선택”을 넘어 여러 조건을 조합해 검색할 수 있게 하면서, DNA 아카이브의 검색 가능성을 한 단계 확장했습니다.

5

검색·연산 확장

최근 흥미로운 흐름은 DNA 저장이 “보관”을 넘어 검색이나 간단한 연산까지 포함하려는 방향으로 확장된다는 점입니다. 전통적으로는 DNA를 시퀀싱해 디지털로 완전히 되돌린 뒤에야 검색을 했습니다. 하지만 이 과정은 시간이 오래 걸리고, I/O 비용도 큼니다. 그래서 DNA 상태에서 바로 유사도 검색이나 필터링을 수행하고, 필요한 결과만 읽어오는 “인-메모리” 접근이 연구되고 있습니다. 예를 들어 Bee 등은 DNA의 상보 결합을 이용해 유사한 데이터만 선택적으로 찾아내는 방식의 가능성을 보여주었고, Nat. Nanotechnol. 2024 연구는 저장된 DNA 위에서 제한적인 계산을 수행하는 방향을 구체화했습니다. 물론 이 분야는 아직 초기 단계이고, 분자 반응의 확률적 잡음, 오프타겟 결합, 조건 민감성 등 해결할 문제가 많습니다. 하지만 초거대 데이터베이스에서 “전부 읽고 계산”하는 방식이 점점 부담스러워지는 상황에서, 이 흐름은 장기적으로 의미 있는 확장 방향으로 평가됩니다.

6

상용화와 국내 시사점

SNIA는 DDS가 아직 상용화 초기 단계이며 해결해야 할 과제가 많다고 전제하면서도, 핵심 구성요소(쓰기-보관-회수-읽기)가 end-to-end 시스템으로 연결될 수 있다는 점이 이미 여러 연구로 확인되었고, 생명과학과 산업의 발전이 DNA 기술을 계속 밀어올릴 것이라고 정리합니다. 또 상용화에서 매우 중요한 요소로 표준화를 강조합니다. 즉, 기술이 좋아도 회사마다 방식이 다르면 산업 생태계가 커지기 어렵기 때문에, 상호운용성과 평가 기준을 만드는 작업이 중요하다는 것입니다. DNA Data Storage Alliance는 인터페이스, 내구성 평가, 메타데이터 구조, 보안/바이오세이프티, 코덱 요구사항 같은 항목을 표준화 대상으로 논의하고 있습니다.

한국은 이 분야에서 드물게 유리한 조합을 갖고 있습니다. 반도체 미세공정은 고밀도 어레이 제작과 정밀 제어에 강점이 있고, 바이오파운드리 역량은 자동화 공정과 DBTL(설계-제작-시험-학습) 사이클 구축에 강점이 있으며, 소프트웨어 역량은 코덱·복호·검색 알고리즘을 빠르게 고도화할 수 있습니다. 따라서 전략적으로 접근한다면, (1) 합성 칩/어레이(화학+효소 혼합 가능), (2) 자동화 공정(파운드리), (3) 코덱·복호·검색 소프트웨어 스택을 함께 묶어 개발하는 로드맵이 가장 설득력 있습니다. 특히 국제 표준화 흐름과 연결해 성능 지표와 검증 체계를 선점할 수 있다면, “기술 따라가기”를 넘어 “산업 규칙 만들기”로 이동할 여지도 생깁니다.

결론적으로 DNA 데이터 저장은 “DNA에 0과 1을 넣는 기술”이 아니라, 저장의 물리적 한계를 넘기 위해 분자 매체를 공학적으로 다루는 인프라 기술입니다. 지난 10여년이 가능성 입증과 랜덤 액세스 구현에 집중했다면, 최근은 자동화·표준화·검색/연산 통합으로 무게중심이 이동하고 있습니다. 비용과 지연시간이라는 장벽은 여전히 크지만, 콜드 데이터 아카이브라는 명확한 적용처가 존재하고 쓰기 기술 혁신과 공정 통합이 맞물린다면, DNA 저장은 장기적으로 에너지·공간·운영비 구조를 바꾸는 선택지가 될 수 있습니다.

< 참고문헌 >

- 1 SNIA. DNA Data Storage Technology Review. Technical White Paper, Version 1.0 (30-June-2025).
- 2 International Data Corporation (IDC) (sponsored by Seagate). Data Age 2025: The Digitization of the World – From Edge to Core. IDC White Paper (2018).
- 3 Ceze, L., Nivala, J., & Strauss, K. Molecular digital data storage using DNA. *Nature Reviews Genetics*, 20(8), 456–466 (2019). DOI: 10.1038/s41576-019-0125-3
- 4 Church, G. M., Gao, Y., & Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science*, 337(6102), 1628 (2012). DOI: 10.1126/science.1226355
- 5 Goldman, N., Bertone, P., Chen, S., et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494(7435), 77–80 (2013). DOI: 10.1038/nature11875
- 6 Bornholt, J., Lopez, R., Carmean, D. M., et al. A DNA-based archival storage system. In: *Proc. ASPLOS '16*, pp. 637–649 (2016). DOI: 10.1145/2872362.2872397
- 7 Erlich, Y., & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science*, 355(6328), 950–954 (2017). DOI: 10.1126/science.aaj2038
- 8 Organick, L., Ang, S. D., Chen, Y.-J., et al. Random access in large-scale DNA data storage. *Nature Biotechnology*, 36(3), 242–248 (2018). DOI: 10.1038/nbt.4079
- 9 Takahashi, C. N., Nguyen, B. H., Strauss, K., & Ceze, L. Demonstration of End-to-End Automation of DNA Data Storage. *Scientific Reports*, 9(1), 4998 (2019). DOI: 10.1038/s41598-019-41228-8
- 10 Heckel, R., Mikutis, G., & Grass, R. N. A characterization of the DNA data storage channel. *Scientific Reports*, 9(1), 9663 (2019). DOI: 10.1038/s41598-019-45832-6
- 11 Newman, S., Stephenson, A., Willsey, M., Nguyen, B., Takahashi, C. N., Strauss, K., & Ceze, L. High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nature Communications*, 10(1), 1706 (2019). DOI: 10.1038/s41467-019-09517-y
- 12 Antkowiak, P. L., Lietard, J., Darestani, M. Z., et al. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nature Communications*, 11(1), 5345 (2020). DOI: 10.1038/s41467-020-19148-3
- 13 Organick, L., Chen, Y.-J., Ang, S. D., et al. Probing the physical limits of reliable DNA data retrieval. *Nature Communications*, 11(1), 616 (2020). DOI: 10.1038/s41467-020-14319-8

- 14 Tabatabaei, S. K., Wang, H., Athreya, N. B., et al. DNA punch cards: storing data on native DNA sequences via topological modifications. *Nature Communications*, 11(1), 1742 (2020). DOI: 10.1038/s41467-020-15588-z
- 15 Banal, J. L., Shepherd, T. R., Berleant, J., et al. Random access DNA memory using Boolean search in an archival file storage system. *Nature Materials*, 20(9), 1272–1280 (2021). DOI: 10.1038/s41563-021-01021-3
- 16 Bee, C., Chen, Y.-J., Zielinski, D., et al. Molecular-level similarity search brings computing to DNA data storage. *Nature Communications*, 12(1), 4764 (2021). DOI: 10.1038/s41467-021-25093-9
- 17 Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie International Edition*, 54(8), 2552–2555 (2015). DOI: 10.1002/anie.201411378
- 18 Lee, H. H., Kalhor, R., Goela, N., Bolot, J., & Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nature Communications*, 10(1), 2383 (2019). DOI: 10.1038/s41467-019-10258-1
- 19 Palluk, S., Arlow, D. H., de Rond, T., et al. De novo DNA synthesis using polymerase-nucleotide conjugates. *Nature Biotechnology*, 36(7), 645–650 (2018). DOI: 10.1038/nbt.4173
- 20 Doricchi, A., Valsesia, D., Baffo, S., et al. Emerging Approaches to DNA Data Storage: Challenges and Prospects. *ACS Nano*, 16(11), 17552–17571 (2022). DOI: 10.1021/acsnano.2c06748
- 21 Lin, K. N., Volkel, K., Cao, C., et al. A primordial DNA store and compute engine. *Nature Nanotechnology*, 19, 1654–1664 (2024). DOI: 10.1038/s41565-024-01771-6
- 22 Landsman, D., & Strauss, K. The DNA Data Storage Model. *Computer*, 56(7), 78–85 (2023). DOI: 10.1109/MC.2023.3272188
- 23 Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M., & Hughes, W. L. Nucleic acid memory. *Nature Materials*, 15(4), 366–370 (2016). DOI: 10.1038/nmat4594

발행일 | 2025년 12월

발행처 | 대전광역시 유성구 과학로 125, KOBIC동 1323호, 합성생물학발전협의회(KSBA) 사무국

전 화 | 042-860-4371~81

이메일 | wogml919@kribb.re.kr, kmyang@kribb.re.kr