DNA 저장 장치에서 길쌈 부호를 이용한 삽입 삭제 오류 정정에 관한 고찰

정재호, 김재원*, 노종선

서울대학교, *경상국립대학교

jhjeong0702@snu.ac.kr, *jaewon07.kim@gnu.ac.kr, jsno@snu.ac.kr

A Study on the Insertion and Deletion Error Correction Using Convolutional Code in DNA Storage System

Jaeho Jeong, Jae-Won Kim*, and Jong-Seon No

Seoul National University, *Gyeongsang National University

요 약

본 논문은 DNA 저장 장치에서 발생하는 삽입 오류와 삭제 오류에 대응하기 위해 길쌈 부호를 활용하였던 최신 연구들을 살펴본다. 특히, 각 연구의 세부 지표들과 함께 어떠한 방식으로 삽입 오류와 삭제 오류에 대응하였는지 간단히 알아보고, 이러한 연구들의 차후 발전 가능성에 대해 알아본다.

Ⅰ. 서 론

차세대 저장 매체로 연구되고 있는 DNA 저장 장치(DNA storage system)[1]에서는 생물학적, 화학적 요인들로 인하여 여러 가지 종류의 오류들이 발생할 수 있다. 이 중에서 대체 오류(substitution error)는 일반적인 오류 정정 부호(error correcting code)나 시퀀스 집단화(sequence clustering) 방식으로 대응이 가능하지만, 삽입 오류(insertion error)나 삭제 오류(deletion error)가 발생한 경우에는 해결하기 어렵다는 단점이 있다. 또한, DNA 시퀀스(sequence)를 복원해내는 과정에서 단일 군집(cluster-size-1) 시퀀스의 경우에는 대체 오류, 삽입 오류, 삭제 오류 모두 발견해내기 어렵고, 자칫하면 다른 부호 워드(codeword)로 정정되어 복호화(decoding) 과정에서 치명적인 영향을 끼칠 수 있게 된다.

지금까지의 DNA 저장 장치 연구에서는 삽입 오류와 삭제 오류가 동시에 발생한 시퀀스의 경우 그림 1과 같이 하나의 커다란 버스트 오류(busrt error)로 간주한 뒤 여러 대체 오류들을 정정하는 방식으로 진행되어 왔으나, 이러한 경우 해당 데이터에 포함되어 있는 부호의 오류 정정 한계를 넘어가는 경우가 종종 발생한다. 이에 단일 군집 시퀀스에서도 활용이 가능하고 삽입/삭제 오류에 대응할 수 있는 오류 정정 부호를 설계하기 위해 DNA 저장 장치에 길쌈 부호(convolutional code)를 적용하는 연구들이 최근에 진행되었다. 본 연구에서는 길쌈 부호를 활용하여 삽입/삭제 오류를 정정하기 위해 진행되었던 여러 최신 연구들에 대해 알아보고, 이러한 연구들의 차후 발전 가능성에 대해 알아본다.

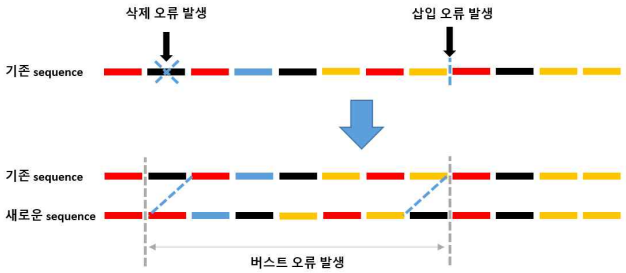

그림 1. 삽입/삭제 오류 발생 시퀀스가 버스트 오류로 처리된 예시

Ⅱ. 본 론

일반적인 통신 시스템에서는 대체 오류를 정정하기 위한 연구가 주로 이루어지지만, DNA 저장 장치에서 발생 가능한 삽입 오류와 삭제 오류에 대응하기 위해서는 새로운 접근 방법이 필요하다. 길쌈 부호 역시 삽입 오류와 삭제 오류를 정정해주지는 않지만, 시퀀스 관점에서 높은 확률을 갖는 비트(bit)들의 수열을 복호화 과정에서 산출해 낸다. 이를 활용하여 시퀀스의 확률이 너무 낮게 복호화가 되면 시퀀스들 내에서 삽입/삭제 오류가 발생했을 가능성을 고려하여 시퀀스의 위치를 좌우로 움직여 복호를 진행하는 방식으로 삽입/삭제 오류를 잡기 위한 연구들이 일부 진행되었고, 또한 삽입/삭제 오류가 발생하면 생길 수 있는 상태들을 상태도(state diagram)에 추가하여 은닉 마르코프 모델(Hidden Markov Model)처럼 추론한 연구도 진행되었다[2]. 이렇게 삽입/삭제 오류를 정정하기 위한 연구들 중 DNA 저장 장치에 실제 실험으로 접목시킨 연구들의 최신 동향에 대해 살펴보도록 한다.

먼저, Chandak et al. 연구진의 논문[3]에서는 길쌈 부호와 RS 부호(Reed-Solomon code, 혹은 RS code)를 이용하여 데이터를 합성하였으며, 길쌈 부호를 삽입/삭제 오류를 고치는데 직접적으로 활용하지는 않았지만, 이를 Nanopore sequencing의 미가공 데이터(raw data)인 전류 흐름의 결과값들과 엮어서 시퀀싱 실험의 결과물을 내놓는 basecalling을 개선시키는 데에 활용하였다. 특히, 길쌈 부호로 만들어질 수 있는 상태도를 머신 러닝의 여러 계층으로 이루어진 신경망으로 활용하여 Nanopore의 머신 러닝 기반 basecalling 알고리즘과 연계하였으며, 이는 추후 Lau et al. 연구진의 연구[4]로까지 발전하였다.

다음으로, Press et al. 연구진의 논문[5]에서도 데이터를 합성하는 데에 길쌈 부호와 RS 부호를 이용하였으며, 부호화 과정에서 길쌈 부호에 해시 함수(Hash function)까지 활용하여 auto-key 암호문을 만드는 것과 비슷한 과정을 거치게 된다. 이후 복호화 과정에서는 매 단계마다 각 시퀀스의 동기화(synchronization) 위치를 비트별로 -1, 0, 1만큼 이동한 것만큼 삽입/삭제 오류 발생 여부를 추론하여 가장 확률 높은 경로(path)를 트리로

생성한 뒤, 스택 알고리즘(stack algorithm)으로 경로를 들고 다니면서 길찾기 알고리즘과 비슷한 방식으로 시퀀스를 추론하게 된다.

마지막으로, Welzel et al. 연구진의 논문[6]에서는 길쌈 부호가 아닌 산술 부호(arithmetic code)와 해시 함수를 사용하였지만, 각 위치별로 0 또는 1을 넣거나 빼면서 순환 중복 검사(Cyclic Redundancy Check, 혹은 CRC)가 통과하는지를 찾으면서 삽입/삭제 오류에 대처하였다. 다음 표1은 위에서 말한 연구들의 세부 지표를 표로 정리한 것이다.

표1. DNA 저장 장치에서 실제 실험을 통해 삽입/삭제 오류 정정 방법을 제시했던 연구들의 세부 지표

	Inner code	Code rate	Outer code
Chandak et al. [3]	길쌈 부호	0.875	RS 부호
Lau et al. [4]	길쌈 부호	0.833	RS 부호
Press et al. [5]	길쌈 부호 + 해시 함수	0.75	RS 부호
Welzel et al. [6]	산술 부호 + 해시 함수	0.5	파운틴 부호

이외에도 DNA 저장 장치에서 Nanopore sequencing을 활용하였을 때의 채널 환경을 가정하여, 길쌈 부호의 상태도를 만들 때 삽입/삭제 오류 상태까지 branch metric을 만들어서 활용하여 시뮬레이션을 통해 해당 채널 모델링의 우수성을 입증한 연구[7]도 진행되었다. 다만 아직까지는 길쌈 부호의 복호화 과정에서 계산 복잡도가 매우 큰 편이기에 이러한 복잡도를 줄이는 연구 역시 필수적으로 동반되어야 할 것이다.

Ⅲ. 결 론

본 논문에서는 일반적인 통신 시스템에서는 자주 발생하지 않지만, DNA 저장 장치라는 특수한 환경에서 발생하는 삽입 오류와 삭제 오류에 대응하기 위해 길쌈 부호를 이용하는 여러 연구들에 대해 알아보았다. 하지만 이렇게 삽입 오류와 삭제 오류가 발생한 시퀀스를 하나의 커다란 오류로 생각을 해본다면 버스트 오류나 심볼(symbol) 오류가 발생한 상황으로도 생각할 수 있고, 이는 다른 메모리나 인터커넥트(interconnect) 시스템 등 여러 연구 분야에서도 충분히 적용 가능한 아이디어라고 볼 수 있다. 이러한 대응 방식들을 통해 새로운 부호 설계에 활용한다면 다른 통신 시스템에서도 적용이 가능한 오류 정정 부호 연구에 큰 도움이 될 것으로 기대된다.

ACKNOWLEDGEMENT

본 연구는 삼성전자의 지원(과제번호: MEM210728_0001)과 한국연구재단을 통해 미래창조과학부의 미래유망융합기술 파이오니어사업(과제번호: 2022M3C1A3081366)으로부터 지원을 받아 수행된 결과임.

참 고 문 헌

[1] 정재호(Jaeho Jeong), 노종선(Jong-Seon No), and 박호성(Hosung Park). "파운틴 코드를 이용한 DNA 저장 장치에서의 효율적인 복호화를 위한 기법," *한국통신학회 학술대회논문집* (2021), 299-300.

[2] Mohamed F. Mansour and Ahmed H. Tewfik, "Convolutional decoding in the presence of synchronization errors," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 2, pp. 218-227, 2010.

[3] Shubham Chandak, Joachim Neu, Kedar Tatwawadi, Jay Mardia, Billy Lau, Matthew Kubit, Reyna Hulett, Peter Griffin, Mary Wootters, Tsachy Weissman, and Hanlee Ji, "Overcoming high nanopore basecaller error rates for DNA storage via basecaller-decoder integration and convolutional codes," in *2020 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 8822-8826, 2020.

[4] Billy Lau, Shubham Chandak, Sharmili Roy, Kedar Tatwawadi, Mary Wootters, Tsachy Weissman, and Hanlee P. Ji, "Magnetic DNA random access memory with nanopore readouts and exponentially-scaled combinatorial addressing," *Scientific Reports*, vol. 13, no. 1, pp. 8514, 2023.

[5] William H. Press, John A. Hawkins, Stephen K. Jones Jr, Jeffrey M. Schaub, and Ilya J. Finkelstein, "HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints," in *Proceedings of the National Academy of Sciences*, vol. 117, no. 31, pp. 18489-18496, 2020.

[6] Marius Welzel, Peter Michael Schwarz, Hannah F. Lochel, Tolganay Kabdullayeva, Sandra Clemens, Anke Becker, Bernd Freisleben, and Dominik Heider, "DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage," *Nature Communications*, vol. 14, no. 1, pp. 628, 2023.

[7] Belaid Hamoum and Elsa Dupraz, "Channel model and decoder with memory for DNA data storage with nanopore sequencing," *IEEE Access*, vol. 11, pp. 52075-52087, 2023.