실시간 리드 필터링 및 클러스터링을 통한 DNA 저장장치 판독 성능 개선

박지연, 박호성*

전남대학교 지능전자컴퓨터공학과

wldus8677@jnu.ac.kr, *hpark1@jnu.ac.kr

Real-time Read Filtering and Clustering for Improved Readout in DNA Data Storage

Jiyeon Park, Hosung Park*

Dept. of Intelligent Electronic and Computer Engineering, Chonnam National Univ.

요 약

본 논문에서는 DNA 저장장치의 실시간 판독 성능 향상을 위해 복호 결과와 리드 간의 불일치 정보를 활용한 저신뢰도 리드 필터링 및 해밍 거리 기반 재클러스터링 기법을 제안한다. 실험 결과, 제안 방법은 기존 방식 대비 최대 4.34%의 판독 비용을 절감하였으며, 오류가 많은 환경에서도 최대 46.37% 낮은 프레임 오류율을 달성하였다. 또한 폐기 리드 분석을 통해 삽입·삭제 오류가 실시간 판독 환경에서 위상 불일치를 유발하는 주요 원인임을 확인하여 향후 특화된 오류 정정 기법과의 결합 가능성을 제시하였다.

Ⅰ. 서 론

DNA 저장장치는 디지털 데이터를 올리고(oligonucleotide, oligo) 형태의 합성 DNA 분자에 저장하는 차세대 저장장치로 주목받고 있다 [1]. 그러나 합성, 증폭 및 시퀀싱 등에 요구되는 화학 실험으로 인해 치환, 삭제, 삽입 오류가 발생하며, 이를 방지하기 위해 LDPC(low density parity check) 부호 같은 오류 정정 기법이 활용되고 있다 [2].

DNA 저장장치의 데이터 판독 과정은 증폭된 올리고 분자로부터 리드(read)라는 염기서열을 생성하는 시퀀싱과 생성된 리드로부터 원본 데이터를 복원하는 복호(decoding)가 포함된다. Illumina 시퀀싱을 비롯한 대부분의 시퀀싱 방식은 서열의 시작부터 끝까지 염기를 하나씩 결정하는 사이클(cycle)이 순차적이며 실시간으로 진행된다. 그러나 기존 시스템에서는 모든 사이클이 종료된 이후에 복호가 수행되므로, 염기 위치마다 올리고 간에 부호가 적용되어 사이클 단위의 복호가 가능함에도 불구하고 중간 복호 결과를 활용할 수 없어 비효율적이다.

한편, DNA 저장장치는 올리고의 순서가 보존되지 않고 증폭 과정에서 다수의 복제본이 생성된다. 올리고를 식별하기 위해 각 올리고에 인덱스(index)를 부여하고, 동일 인덱스를 갖는 리드들을 클러스터링(clustering)하는 구조를 사용할 수 있다. 이러한 특성과 사이클 단위 복호가 가능한 오류 정정 부호의 특징을 이용한다면 복호 결과와 리드 간의 일치도를 실시간으로 평가하여 낮은 신뢰도의 리드를 조기에 배제할 수 있다.

본 논문에서는 시퀀싱 사이클마다 염기를 결정하는 베이스 콜링(base-calling)과 복호를 동시에 수행하며, 복호 결과와 리드 간의 불일치 정보를 누적하여 저신뢰도 리드를 실시간으로 필터링하는 방법을 제안한다. 또한, 주기적으로 저신뢰도 클러스터에 대해 해밍 거리 (hamming distance) 기반 클러스터링을 수행하여 클러스터 신뢰도를 보완한다. 제안한 방법을 통해 데이터 복원 성능을 개선함으로써 비교군 대비 최대 4.34%의 판독 비용을 절감하였다.

Ⅱ. 제안하는 방법

2.1 인덱스 기반 초기 클러스터링

본 논문에서 고려하는 올리고는 인덱스 영역과 그 이후의 데이터 영역으로 구성된다. 이에 따라 초반 시퀀싱 사이클이 인덱스 길이만큼 진행된 후에 인덱스 염기가 결정되며, 인덱스에 부여된 오류 검출 부호를 이용해 인덱스의 유효성을 판단할 수 있다. 유효한 인덱스를 가진 리드들은 사전에 알려진 순서에 따라 정렬되어 인덱스별로 클러스터를 형성하며, 각 클러스터는 하나의 올리고 위치에 대응된다. 이후, 데이터 영역의 시퀀싱 및 복호가 진행되며, 오류 인덱스를 가진 리드는 재활용 전까지 베이스 콜링만 수행하고 복호에는 사용하지 않는다.

2.2 실시간 저신뢰도 리드 필터링

각 사이클에서 복호가 성공하면 해당 염기의 정답을 알 수 있으며, 이를 기반으로 클러스터 내의 리드와 정답 간의 일치도를 평가한다. 리드 i에 대해 첫 불일치 발생 시점을 $t_{diff}^{(i)}$라 할 때 수식 1과 같이 임계값 τ 이상의 리드는 복호를 위한 LLR(log-likelihood) 계산에서 제외한다, t는 현재 사이클이며 M은 정답과 불일치하는 염기의 개수이다. 또한, 누적 불일치 개수가 c 이상인 경우 해당 리드를 클러스터에서 완전히 폐기하여 향후 연산에 사용하지 않는다.

$$\frac{M(t_{diff}^{(i)}, t)}{t - t_{diff}^{(i)} + 1} \geq \tau \quad \text{(수식 1)}$$

2.3 주기적 해밍 거리 기반 재클러스터링

클러스터 크기 $|cluster|$는 클러스터에 포함된 리드의 개수로 정의되며, 크기가 작을수록 오류의 영향을 크게 받는다. 따라서 클러스터 신뢰도 $R_{cluster}$를 수식 2와 같이 정의하고, 신뢰도가 r 이하인 클러스터에 대해서만 해밍 거리 기반 재클러스터링을 수행한다. 해당 재클러스터링은 주기 p마다 실행되며, 클러스터의 정답과 해밍 거리가 d 이하인 오류 인덱스 리드를 해당 클러스터에 병합한다. 이때 해밍 거리는 첫 사이클부터 현재 사이클까지의 염기서열을 기준으로 계산한다.

$$R_{cluster} = \begin{cases} \dfrac{\sum_{i \in C}(t - M(t_{diff}^{(i)}, t))}{t \times |cluster|}, & |cluster| > 0 \\ 0, & |cluster| = 0 \end{cases} \quad \text{(수식 2)}$$

Ⅲ. 모의실험

Baseline은 Illumina의 베이스 콜링 프로그램인 Bustard [3]로부터 생성된 리드에 인덱스 기반 클러스터링 및 복호를 진행한 결과이다. 제안 방법의 성능을 검증하기 위해 Baseline에 추가 과정을 수행하는 두 시나리오를 구성하였다. Proposed-F는 2.2절의 저신뢰도 리드 필터링만 적용하며, Proposed-FC는 필터링뿐만 아니라 2.3절의 재클러스터링도 적용한다.

올리고의 전체 길이는 151 nt (nucleotide)이며, 염기 - 비트 매핑은 A:00, C:01, G:10, T:11로 정의하였다. 초반 16nt는 CRC(cyclic redundancy check)-15가 적용된 인덱스 영역이며, 이후 135 nt 데이터 염기 영역에는 LDPC(73728, 63621) 부호가 사이클 단위로 적용되었다. 실험 파라미터들은 $\tau = 0.5, c = 4, r = 0.9, p = 40, d = 3$로 설정하였다.

그림 1은 무작위 추출된 리드 개수에 따른 프레임 오류율을 보여준다. Baseline의 경우, 460,000개의 DNA 분자를 사용했을 때 오류 없는 완전한 데이터 복원이 가능하였다. 반면 Proposed-F는 450,000개, Proposed-FC는 440,000개의 더 적은 리드만으로 완전 복원이 가능하여, 각각 2.17%와 4.34%의 판독 비용 절감을 달성하였다. 또한, 가장 적은 430,000개의 리드만 사용한 경우에도 Proposed-F와 Proposed-FC는 Baseline 대비 각각 약 25.7%와 46.37% 낮은 프레임 오류율을 보였다.

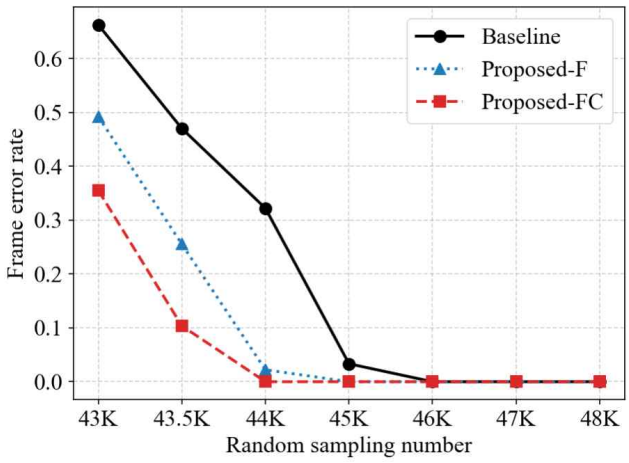

그림 1. 무작위 추출된 리드 개수에 따른 오류 정정 성능 비교

Proposed-FC가 완벽한 데이터 복원을 달성한 440,000개의 리드 상황에서, 72,000개의 전체 클러스터와 113,430개의 오류 인덱스 리드에 대해 3번의 재클러스터링이 수행되었다. 수행 시점마다 3.31%, 2.92%, 2.95%의 클러스터가 저신뢰도로 판단되었고, 158, 10, 3개의 오류 인덱스 리드가 클러스터로 편입되었다.

추가적으로, 폐기된 리드의 활용성 분석을 위해 440,000개의 리드 상황에서 10,765개의 폐기 리드에 대해 모든 올리고와의 최소 편집 거리(edit distance)를 계산하였다. 그림 2는 전체 폐기된 리드 중 편집 거리 20 이하인 리드를 대상으로, 각 편집 거리별로 리드에 포함된 치환, 삭제, 삽입 오류의 비율과 해당 편집 거리를 갖는 리드의 개수 비율을 나타낸 것이다.

폐기된 전체 리드 중 48.88%는 편집 거리 3 이하였으며, 이 중 90.48%는 삽입과 삭제 오류만으로 구성되어 있었다. 이는 적은 삽입·삭제 오류만으로도 실시간 판독 환경에서는 연속적인 불일치를 유발하여 위상 차이가 발생할 수 있음을 의미한다.

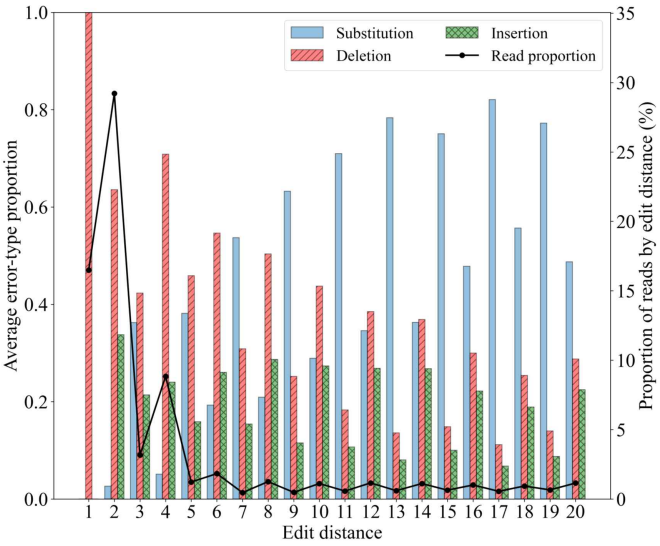

그림 2. 편집 거리별 오류 및 리드 개수의 분포

Ⅳ. 결론

본 논문에서는 DNA 저장장치의 실시간 데이터 판독 환경에서 염기 결정과 복호를 동시에 수행하고, 복호 결과와 리드 간의 불일치 정보를 누적하여 저신뢰도 리드를 조기에 필터링하는 방법을 제안하였다. 또한, 클러스터 신뢰도가 낮은 경우에 한해 해밍 거리 기반 재클러스터링을 수행함으로써 클러스터의 신뢰도를 보완하였다. 실험 결과, 제안한 방법들은 기존 방식 대비 판독에 요구되는 리드 수를 줄여 판독 비용을 2.17%-4.34% 절감하였다. 또한, 폐기된 리드 대부분이 삽입 및 삭제 오류로 인해 폐기되는 것을 분석하여 이에 특화된 검출 및 정정 기법을 적용할 시 추가적인 성능 향상이 기대됨을 확인하였다. 다만, 소수의 리드는 세 오류가 혼합되어 존재하므로, 단일 오류 가정에 기반한 정정 기법의 오동작을 방지하기 위해 오류 유형을 식별하거나 분리할 수 있는 방안이 필요하다.

ACKNOWLEDGMENT

이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원 (No. 2022M3C1A3090857, No. RS-2024-00410005) 및 정보통신기획평가원-지역지능화혁신인재양성사업(IITP-2026-RS-2022-00156287, 10%)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] Matange, K., Tuck, J.M. and Keung, A.J. “DNA stability: a central design consideration for DNA data storage systems.” Nature Communications, 12(1): 1358, 2021.

[2] S. J. Park, J. H. Jeong, S. H. Kim, A. No, J. S. No, and H. Park. “Reducing cost in DNA-based data storage by sequence analysis-aided soft information decoding of variable-length reads.” Bioinformatics, 39(9), 2023.

[3] Cacho, A., Smirnova, E., Huzurbazar, S., and Cui, X..“A comparison of base-calling algorithms for illumina sequencing technology.” Briefings in bioinformatics, 17(5), pp.786-795, 2016.