DNA 저장장치 실험 재현을 위한 오류 시뮬레이터 연구

박지연, 박호성

전남대학교 지능전자컴퓨터공학과

wldus8677@gmail.com, hpark1@jnu.ac.kr

A Study on the Error Simulator to Reproduce DNA storage Experiments

Jiyeon Park, Hosung Park

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

요 약

차세대 저장 매체로 주목받고 있는 DNA 저장장치는 생화학적 과정에서 발생하는 오류를 해결하기 위해 오류 정정 부호 등의 오류 정정 기법이 필수적으로 사용된다. 그러나 DNA 저장장치에 최적화된 오류 정정 개발을 위해서는 생화학 실험을 반복해야 하므로 많은 시간과 비용이 소요된다. 이를 보완하기 위해 실제 DNA 저장장치 실험 환경을 컴퓨터상으로 모사하는 오류 시뮬레이터 연구가 진행됐으나 이는 내부 실험들을 개별적으로 고려하므로 DNA 저장장치의 오류 정정을 최적화하고자 하는 목적에서는 오히려 복잡하다. 이에 DNA 저장장치의 실험들을 하나의 채널로 고려하고 통합된 채널의 오류를 추정하여 새로운 데이터가 입력되더라도 주어진 DNA 저장장치 오류를 재현할 수 있는 오류 시뮬레이터를 제안한다. 제안하는 시뮬레이터로 생성된 데이터와 실제 실험 데이터의 복호 성능을 비교하여 제안된 방법이 안정적인 오류 재현이 가능함을 보인다.

Ⅰ. 서 론

저전력, 높은 저장 밀도 등으로 차세대 저장 매체로 주목받고 있는 DNA 저장장치는 그림 1과 같이 합성 (synthesis), 중합 효소 연쇄반응 (polymerase chain reaction: PCR), 시퀀싱 (sequencing) 등 시험관 내에서 (in vitro) 진행되는 생화학 실험이 요구된다 [1]. 완벽한 통제가 어려운 실험으로 인해 각 실험에서는 원인 규명이 어려운 오류들이 발생한다.

오류에 대응하기 위해 DNA 저장장치는 오류 정정 부호와 같은 오류 정정 기법의 사용을 필수적으로 채택하고 있다 [2]. 그러나 DNA 저장장치에 최적화된 오류 정정 기법을 개발하기 위해서는 오류 정정 기법의 변수가 변경될 때마다 모든 실험 과정들을 반복해야 하므로 많은 시간과 비용이 소요된다는 단점이 존재한다.

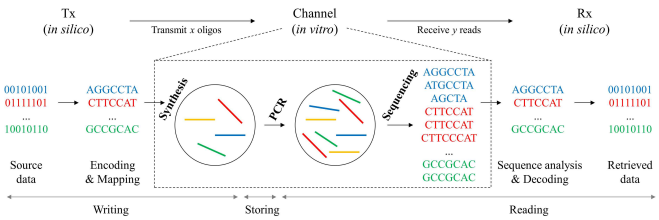

그림 1. DNA 저장장치를 통한 데이터 복원 과정

이를 보완하기 위해 DNA 저장장치 실험의 오류 환경을 컴퓨터 시뮬레이션 실험으로 (in silico) 모사하는 오류 시뮬레이터 연구가 진행되어왔다. 기존 연구들은 합성, PCR 등의 실험 별로 사전에 경험적으로 분석된 오류 특성을 사용하여 DNA 저장장치의 오류를 재현했다 [3, 4].

그러나 생화학 실험들을 하나의 통신 채널처럼 여기는 DNA 저장장치에서 실험마다 독립적으로 오류를 최적화하는 것은 오류 정정 방식을 연구하는 관점에서는 오히려 복잡한 작업이다. 또한, 오류 정정 전문가가 생화학 실험의 전문 지식을 숙지해야 하는 단점도 존재한다.

이에 임의의 DNA 저장장치가 주어졌을 때 실험의 전체 과정을 하나의 통합된 오류 채널로 고려하고, 채널 오류를 분석하는 오류 시뮬레이터를 제안한다. 제안하는 시뮬레이터는 채널의 입출력을 비교하여 오류의 특성을 파악하고 이를 기반으로 새로운 DNA 서열에 해당 DNA 저장장치의 오류를 재현할 수 있다. 오류 재현의 정확도를 판단하기 위해 실제 DNA 저장장치 실험으로 생성된 DNA 서열과 시뮬레이터로 생성된 가상 DNA 서열 간의 복호 성능을 비교하며, 제안한 오류 시뮬레이터가 실제 실험에 매우 근접한 오류 특성을 재현하는 것을 확인할 수 있다.

Ⅱ. 제안하는 방법

제안하는 오류 시뮬레이터는 그림 2와 같이 저장하고자 하는 디지털 데이터가 DNA 서열로 변환된 올리고 (oligo)와 이를 실제 DNA 저장장치를 통해 읽어낸 DNA 서열인 리드 (read)를 입력받는다. 리드는 DNA 저장장치 실험이 진행되면서 올리고로부터 증폭되고 오류가 발생한 데이터이다.

DNA 저장장치 실험상 오류를 추정하기 (error estimation) 위해 컴퓨터상에서 올리고와 리드를 편집거리로 [5] 비교한다. 이를 통해 올리고마다 리드로 증폭된 비율, 각 리드 내에서 오류가 발생한 염기 위치, 발생한 오류의 종류를 추정하여 오류 프로필 (profile)을 생성한다. 오류 종류는 기존 염기가 다른 염기로 변하는 대체 오류, 기존 염기가 사라지는 삭제 오류, 새로운 염기가 추가되는 삽입 오류 중 하나이다.

이후 새로운 올리고들을 입력받아 올리고마다 오류 프로필을 기반으로 오류를 재현한다. 새로운 올리고는 주어진 DNA 저장장치의 제한을 따라야 하므로 기존 올리고와 서열 길이 및 개수가 동일하나 사용된 디지털 데이터, 오류 정정 부호의 종류, 염기 순서 등은 다를 수 있다.

오류 재현은 (error reproduction) 다음과 같은 순서로 진행된다. 먼저, 해당 올리고를 증폭된 비율만큼 복사한다. 복사된 올리고들은 모두 다른 리드의 특성을 따르게 된다. 복사된 올리고에 대응되는 리드 오류 염기 위치들을 서열의 시작부터 끝까지 정렬한다. 정렬된 순서에 따라 올리고 데이터에 해당 위치에서 발생한 오류를 주입한다. 대체 오류일 경우, 해당 위치의 올리고 염기를 해당 염기가 아닌 다른 염기로 무작위로 교체한다.

삽입 오류의 경우, 현재 위치의 앞에 A, C, G, T 네 종류의 염기 중 하나를 무작위로 추가한다. 삭제 오류의 경우, 해당 위치에 존재하는 올리고 염기를 삭제한다. 이를 통해 주어진 DNA 저장장치 실험의 오류 특성을 따르는 새로운 리드를 생성할 수 있다.

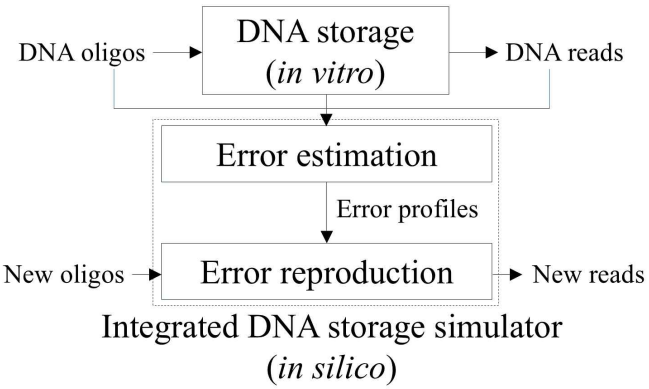

그림 2. 제안하는 DNA 저장장치 오류 시뮬레이터

Ⅲ. 모의실험

제안하는 오류 시뮬레이터의 정확도를 판단하기 위해 실제 DNA 저장장치 실험으로 생성된 리드와 시뮬레이터로 생성된 가상 리드 간의 복호 성능을 비교한다. 사용된 DNA 저장장치는 LT (luby transform) 와 RS (reed solomon) 부호가 각각 외부, 내부 부호로 적용된 152 nt (nucleotide) 길이의 올리고 18000개가 약 1400만개 리드로 증폭된 실험이 진행되었다.

복호는 [2]와 동일한 방법을 사용하였다. 양방향 리드를 하나의 리드로 합친 후, 152 nt 길이를 가지는 리드만을 사용한다. 이후 동일한 리드를 같은 클러스터로 만드는 클러스터링을 수행하며, 클러스터 내에 존재하는 크기가 큰 순서대로 클러스터를 정렬한다. 해당 순서대로 클러스터를 복호기에 입력하여 RS 부호로 오류를 검출한 후, 오류가 검출되지 않은 DNA 서열만을 사용하여 LT 부호의 삭제 (erasure) 복호를 수행한다.

복호 성능을 비교하기 위해 전체 리드 집합에서 무작위로 특정 개수의 리드를 추출한 후 위에 언급된 복호 과정을 수행하는 모의실험을 진행하였다. 실제 DNA 저장장치 실험의 리드로 모의실험한 경우를 Original, 동일한 올리고를 입력하여 시뮬레이터로 생성한 리드로 모의실험한 경우를 Simulation이라 지칭한다. 리드 추출 및 복호는 독립적으로 100번 수행하였으며, 이에 따른 두 모의실험의 복호 성능을 추출 리드 개수에 따라 그림 3에 나타내었다.

모든 경우에서 Original과 Simulation의 성능은 완벽하게 일치하였다. 74000개의 리드를 추출했을 때는 100번의 복호 시도 중 모든 시도가 실패하였으며, 105000개의 리드를 추출했을 때는 100번의 복호 시도가 모두 성공하였다. 오류 시뮬레이션 과정에서 어떤 염기가 다른 염기로 변환 (transformation) 되었다거나 DNA 저장장치 실험들의 오류 분포 변화를 추적하는 등의 복잡한 추정 및 재현을 하지 않고 보다 단순한 편집거리 기반 오류 분석만을 수행했음에도 실제 DNA 저장장치 실험에 근접하는 결과를 보였다는 것을 확인할 수 있다. 이는 결과적으로 제안하는 오류 시뮬레이터가 주어진 DNA 저장장치 실험의 오류 경향을 안정적으로 재현하고 있다는 것을 의미한다.

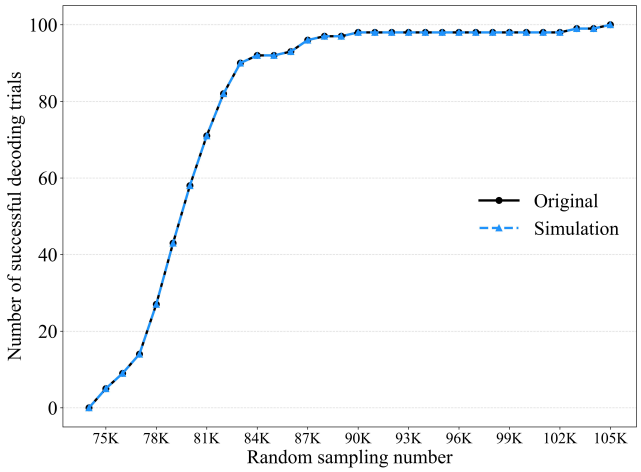

그림 3.무작위 추출된 리드 개수에 따른 복호 성능 비교

Ⅳ. 결론

본 논문에서는 DNA 저장장치의 생화학 실험들로 발생하는 오류를 컴퓨터 시뮬레이션으로 재현하는 오류 시뮬레이터를 제안하였다. DNA 저장장치의 실험들을 하나의 통신 채널로 고려하여 채널의 입출력 데이터를 편집거리로 분석하였고, 오류 종류 및 발생 위치별로 오류를 재현하였다.

제안한 시뮬레이터는 주어진 DNA 저장장치에 대해 기존 방법들보다 간단하게 DNA 저장장치의 오류를 계산하지만, 실제 실험과 굉장히 유사한 오류 경향을 재현할 수 있다. 또한. 내부 실험들에 대한 전문 지식을 요구하지 않으므로 DNA 저장장치를 위한 오류 정정 연구에 쉽게 접근할 수 있다. 향후 제안한 시뮬레이터를 활용하여 DNA 저장장치의 복호 성능 분석 및 오류 정정 성능 개선 연구를 진행할 예정이다.

ACKNOWLEDGMENT

본 연구는 한국연구재단을 통해 미래창조과학부의 미래유망 융합기술 파이오니어사업과 (과제번호-2022M3C1A3090857) 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화 혁신인재양성사업 (IITP-2024-00156287, 20%) 및 한국연구재단의 지원을(No. RS-2024-00410005) 받아 수행된 연구임.

참 고 문 헌

[1] R. Heckel, G. Mikutis, and R. N. Grass, “A characterization of the dna data storage channel,” Scientific reports, vol. 9, no. 1, p. 9663, 2019.

[2] Y. Erlich, D. Zielinski, “DNA Fountain enables a robust and efficient storage architecture”. Science, vol. 355, issue. 6328, pp. 950-954, 2017.

[3] M. Schwarz, M. Welzel, T. Kabdullayeva, A. Becker, B. Freisleben, and D. Heider, “Mesa: Automated assessment of synthetic dna fragments and simulation of dna synthesis, storage, sequencing and pcr errors,” Bioinformatics, vol. 36, no.11, pp. 3322-3326, 2020.

[4] L. Yuan, Z. Xie, Y. Wang, and X. Wang, “Desp: A systematic dna storage error simulation pipeline,” BMC bioinformatics, vol. 23, no. 1, p. 185, 2022.

[5] V. I. Levenshtein et al., “Binary codes capable of correcting deletions, insertions, and reversals,” Soviet physics doklady, vol. 10, no. 8, pp. 707-710, 1966.