DNA 저장장치의 복호를 위한 확장된 서열 활용

박지연, 박호성
전남대학교
wldus8677@jnu.ac.kr, hpark1@jnu.ac.kr

Expanded sequence utilization for decoding of DNA storage

Jiyeon Park, Hosung Park
Chonnam National Univ.

요 약

차세대 저장매체로 주목받고 있는 DNA 저장장치는 신뢰적인 데이터 저장을 위해 오류 정정 부호의 사용이 필수적이나 이로 인해 DNA 저장장치의 데이터 쓰기 비용과 읽기 비용이 상충하는 문제가 발생한다. 본 논문에서는 기존의 DNA 저장장치 과정에서 버려졌던 서열을 추가적으로 복호기의 입력으로 활용함으로써 확장된 서열을 사용하는 방법이 복호 성능을 향상할 수 있음을 보였다. 제안한 방법은 읽기 비용은 약 7.4% 절감하면서 쓰기 비용은 유지할 수 있다는 장점이 존재한다.

Ⅰ. 서 론

스마트 장치의 보급과 IT 서비스의 발전으로 인해 방대한 양의 디지털 데이터가 생성되고 있다. 많은 양의 데이터를 장기간 보관하기 위해 기존 저장매체보다 전력, 안정성 등에서 장점이 있는 DNA 서열에 데이터를 저장하는 DNA 저장장치가 차세대 저장매체로 주목받고 있다.

DNA 저장장치는 크게 데이터 쓰기 과정에 해당하는 합성 (synthesis)과 데이터 읽기 과정인 중합 효소 연쇄반응 (polymerase chain reaction: PCR) 및 시퀀싱 (sequencing) 으로 구성된다. 해당 과정들은 생화학적으로 진행되므로 완벽하게 통제될 수 없기에 모든 과정에서 오류가 발생하게 된다. 그렇기에 신뢰적인 데이터 저장을 위해 오류 정정 부호가 필수적으로 요구된다.

오류 정정 부호는 리던던시 (redundancy) 라는 정보를 추가하여 오류를 견딜 수 있지만, 저장 가능한 정보의 밀도가 낮아진다는 단점이 있다. 리던던시를 적게 사용한다면 정보 밀도를 높일 수 있으나 복호기에서 데이터를 완벽하게 복원하는데 더욱 많은 서열 정보를 요구하게 하므로 읽기 비용이 증가하게 된다[1]. 즉, 리던던시의 수와 비례하는 쓰기 비용과 읽기 비용 간에 상충하는 문제가 발생한다.

본 논문에서는 정보 밀도 및 쓰기 비용이 고정된 상태에서 복호 성능을 개선하여 읽기 비용을 감소하는 것을 목표로 한다. 이에 DNA 저장장치 과정에서 버려지는 서열을 활용하여 복호기가 사용 가능한 서열 정보를 추가로 확보하는 서열 확장 방법을 제안한다. 기존의 서열 분석 방법[2]에 서열 확장을 적용한 결과, 결과적으로 7.4%의 읽기 비용을 절감할 수 있다는 것을 보였다.

Ⅱ. 본론

2.1 실험 환경

본 논문의 합성, PCR을 비롯한 DNA 저장장치 실험은 기본적으로 [3]의 방법을 따른다. 513.6KB의 용량을 가지는 이미지 데이터를 256 비트의 16,050개의 조각으로 나눈 후. LT (luby transform) 부호를 통해 1.12배의 리던던시를 추가하여 총 18,000개의 조각을 생성하였다. 또한, 각 조각에 (38, 36) RS (reed solomon) 부호를 GF(2^8)에서 적용하여 해당 데이터 조각에 오류 유무를 판단할 수 있도록 하였다. 이후. [3]와 동일하게 DNA 서열로의 매핑, 합성, PCR 및 시퀀싱을 진행하였다. 결과적으로 152 nt 길이의 DNA 서열들이 저장되었다.

이때 사용한 일루미나 시퀀싱은 순도 (chastity) 값을 통해 시퀀싱된 서열의 질을 판단하여 버리는 필터링 과정이 존재한다. 이렇게 버려진 서열을 NPF (not passing filter: NPF) 서열이라 부르도록 하며 필터를 통과하여 시퀀싱의 최종 결과로 출력되는 서열을 PF 서열이라 하도록 한다. 결과로 제공되지 않는 NPF 서열까지 분석하기 위해 시퀀싱 중의 원시 데이터를 추출하여 이를 AYB라는 base caller를 통해 온전한 DNA 서열로 읽어내었다. 읽어낸 서열에는 동일한 양의 정방향 (R1) 및 역방향 (R2) 서열이 있으며, R1 PF 서열은 2,746,447개, NPF는 325,527개로 PF 서열이 약 8.44배 더 많이 존재한다.

2.2 서열 분석

DNA 저장장치의 복호를 위해서는 읽어낸 서열을 처리하는 서열 분석 과정이 먼저 수행된다. 본 논문에서는 서열의 품질을 높이기 위해 두 방향 서열을 합하는 병합 (merging) 과정을 수행하였다. 병합 후의 정상 길이를 l이라 할 때, $l = 152$이며 이에 해당하는 병합된 서열의 비율은 PF는 82.43%, NPF는 56.41%를 차지한다.

두 서열의 품질을 추정하기 위해 저장했던 서열과 Levenshtein 거리를 측정하고 오류율을 계산하여 그림 1에 나타내었다. 오류는 염기가 대체되는 치환 (Sub), 존재했던 염기가 사라지는 삭제 (Del), 없던 염기가 생기는 삽입 (Ins) 오류로 분류된다. 전반적으로 NPF 서열의 오류율이 PF 서열보다 높으므로 NPF 서열의 품질이 낮은 것을 확인할 수 있다. 하지만, NPF 서열 중에서도 정상 길이인 152 nt의 NPF 서열은 다른 길이의 서열에 비해 낮은 오류율을 보이므로 이러한 서열은 오류가 없는 정상 서열

이 포함되어있을 가능성이 있다.

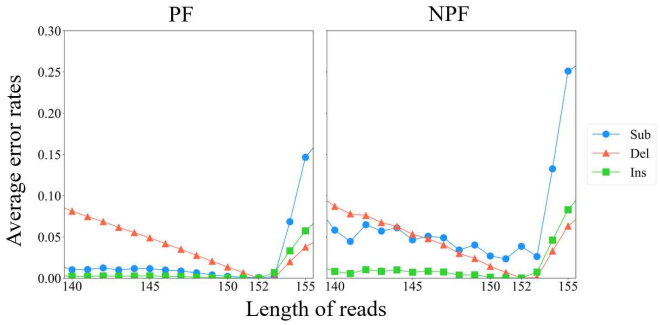

그림 1. 병합된 서열의 길이에 따른 평균 오류율 비교

PF와 NPF 서열에 포함된 정상 서열를 추출하기 위해 오류 정정 부호를 활용할 수 있다. 각 서열은 2.1절에서 언급한 RS 부호를 통해 오류 검출이 가능하며, 이로 인해 오류가 검출되지 않은 정상 서열만을 복호기가 사용할 수 있다. 기존 연구 중, Erlich[2]는 PF 서열과 오류 검출기가 포함된 서열 분석 방법을 사용하여 좋은 복호 성능을 달성한 결과를 보여주었다.

Erlich의 서열 분석 방법은 그림 2와 같다. 먼저 R1, R2 서열들을 병합한 후, 길이가 l인 서열만을 클러스터링에 사용한다. 이때 사용한 클러스터링은 모든 염기가 동일한 서열들을 하나의 클러스터로 만드는 작업이며, 각 클러스터의 크기는 클러스터 내에 존재하는 서열의 개수이다. 클러스터 내에는 같은 서열만 존재하므로 클러스터 내의 임의의 서열을 대표 서열로 결정한다. 클러스터의 크기가 큰 순서로 대표 서열을 정렬한 후, 오류 검출기 및 복호기에 입력한다. 클러스터의 크기가 큰 순서부터 입력하는 것은 서열의 신뢰성을 고려한 것으로, 오류를 포함한 서열이라면 이와 같은 오류 구조를 갖는 서열은 대부분 없기에 오류 서열의 작은 클러스터 크기를 보장한다.

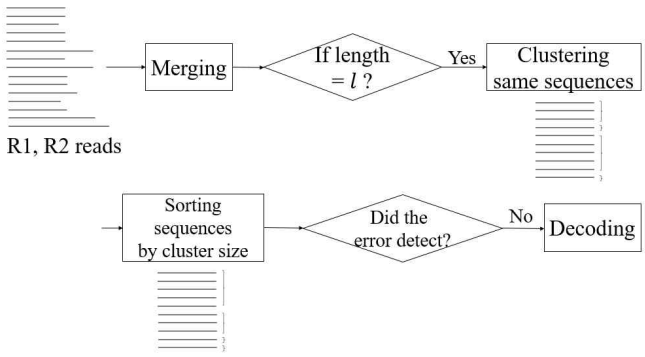

그림 2. Erlich의 서열 분석 방법 ($l = 152$)

2.3 모의실험

본 논문에서 제안하는 방법은 기존에 버려졌던 NPF 서열까지 사용하는 서열 확장이다. 제안하는 방법의 효용성을 확인하기 위해 Erlich의 서열 분석 방법인 Erlich-PF와 제안한 서열 확장을 적용한 Erlich-All의 복호 성능을 비교하도록 한다.

복호 성능은 서열을 무작위 추출한 집합 100개에 대해 복호의 성공 횟수로 측정하였다. 추출된 서열의 개수가 96,000일 때 Erlich-All이 Erlich-PF보다 약 67.28개의 정상 서열을 추가로 확보하였다. 또한, 추출 개수가 100,000일 때 Erlich-All이 모든 집합을 완벽히 복호한 것에 비해 Erlich-PF는 그를 초과한 108,000일 때 성공하였다. 이는 같은 개수의 서열을 사용할 때, NPF 서열에서 확보된 정상 서열들이 서열 부족으로 인해 실패했어야 할 복호를 성공시키는 것과 약 7.4%의 읽기 비용이 감소했다는 것을 의미한다.

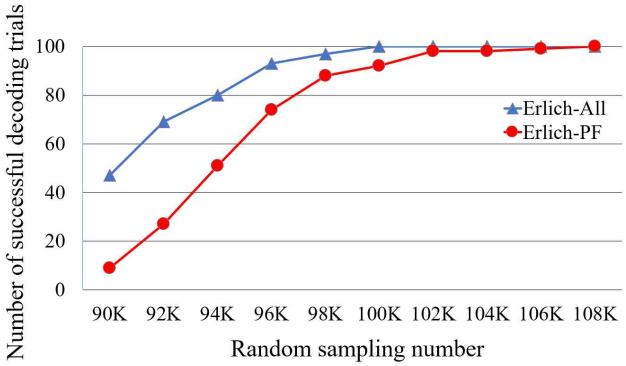

그림 3. 무작위 추출된 서열 개수에 따른 복호 성능 비교

Ⅲ. 결론

본 논문에서는 DNA 저장장치의 쓰기 비용은 유지하면서 읽기 비용의 이득을 얻기 위해 복호 성능을 개선하는 방법을 제안하였다. 해당 방법은 일루미나 시퀀싱 과정에서 순도 필터에 의해 버려지는 NPF 서열을 사용하여 복호기가 사용 가능한 서열을 확장하는 것이다. 오류 검출기를 통해 NPF 서열에 포함된 정상 서열을 추출할 수 있으며, 이는 완벽한 복호를 위해 부족한 서열 정보를 보완하는 역할을 한다.

모의실험을 통해 쓰기 비용이 제한된 상황에서 제안한 방법이 기존의 서열 분석 방법보다 읽기 비용을 약 7.4% 절감할 수 있음을 확인하였다. 이후 연구에서는 서열 분석 방법을 개선하여 정상 길이를 가지지 않은 서열까지 활용하는 서열 확장 연구를 진행할 계획이다.

ACKNOWLEDGMENT

본 논문은 한국연구재단을 통해 미래창조과학부의 미래유망 융합기술 파이오니어사업 (과제번호-2022M3C1A3090857)과 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재 4.0 사업 (IITP-2022-RS-2022-00156385)의 연구결과임

참 고 문 헌

[1] R. Heckel, I. Shomorony, K. Ramchandran and D. N. C. Tse, "Fundamental limits of DNA storage systems", 2017 IEEE International Symposium on Information Theory (ISIT), pp. 3130-3134, 2017.

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

[3] J. Jeong, S. J. Park, J. W. Kim, J. S. No, H. H. Jeon, J. W. Lee, and H. Park, "Cooperative sequence clustering and decoding for DNA storage system with fountain codes", Bioinformatics, vol. 37, issue. 19, pp. 3136-3143, 2021.