DNA 저장장치에서 FASTQ 파일 정보를 활용한 연판정 복호화 기법

김재원*, 정재호, 노종선
*경상국립대학교, 서울대학교

*jaewon07.kim@gnu.ac.kr, jaehoj@ccl.snu.ac.kr, jsno@snu.ac.kr

Soft Decision Decoding Using FASTQ Files for DNA Storage

Jae-Won Kim*, Jaeho Jeong, Jong-Seon No
*Gyeongsang National University, Seoul National University

요 약

본 논문은 DNA 저장장치의 시퀀싱 과정에서 제공되는 FASTQ 파일로부터 염기 확률을 얻고 채널 통계 정보를 바탕으로 log likelihood ratio (LLR)를 계산하여 연판정 복호화를 수행한다. 이로부터 기존 강판정 기반의 복호화 방식 대비 복호화 성능 개선을 시퀀싱 리드 수 관점에서 실험을 통해 검증한다.

Ⅰ. 서 론

최근 빅데이터 기반의 기술들과 산업들이 발전함에 따라 저장해야하는 데이터의 양이 크게 증가하고 있기 때문에 현재 대표적인 저장장치인 hard disk drive (HDD)나 solid state drive (SSD)보다 저장 밀도 관점에서 우수한 차세대 저장장치가 필요한 상황이다. 이에 저장 밀도 관점에서 우수한 차세대 저장장치로서 DNA 저장장치가 활발히 연구되고 있다. DNA 저장장치는 DNA 의 염기서열을 사용자가 원하는 순서대로 합성하여 정보를 저장하는 방식이며 염기의 종류는 4 가지로 adenine, guanine, cytosine, thymine 이고 각각을 편의상 A, G, C, T 로 표현한다. 다른 저장장치와 마찬가지로 DNA 저장장치도 쓰기와 읽기가 존재하는데 쓰기는 DNA 합성을 의미하고 읽기는 DNA 시퀀싱을 의미한다.

DNA 시퀀싱에 앞서 저장된 DNA 의 수를 증가시키기 위해 polymerase chain reaction (PCR)을 진행하며 흔히 자연계에 존재하는 DNA 는 이중나선 구조이지만 DNA 저장장치에서는 single strand 를 사용한다. DNA 시퀀싱 과정은 흔히 양방향으로 진행되며 양방향으로 읽은 정보를 합치는 알고리즘이 필요한데 대표적인 알고리즘이 PEAR 이다 [1]. DNA 저장장치는 생물학적 특성으로 인해 기존 저장장치와는 다른 형태의 오류들이 발생하며 대표적으로 삽입 오류, 삭제 오류, 대체 오류 등이 있다. 더불어 DNA 합성과 시퀀싱 과정에서 homopolymer runlength, GC-content 등의 제약조건에 따라 오류율이 다르기 때문에 이를 고려하여 정보를 저장하는 작업이 필요하다. 이렇듯 기존 저장장치와는 다른 특성을 보이지만 DNA 라는 유기물은 저장 밀도 관점에서 기존 저장장치들보다 수 order 이상 효율적이다.

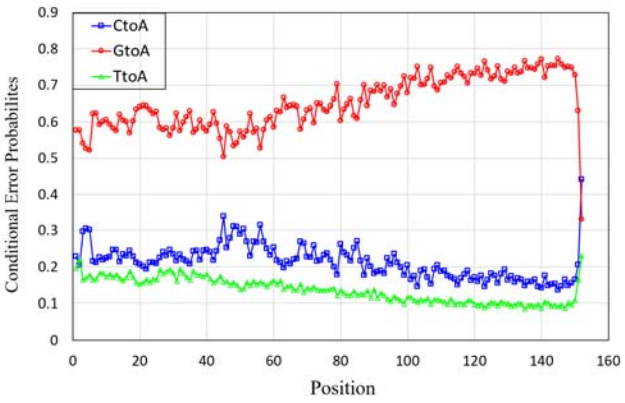
DNA 저장장치에 오류정정부호를 적용한 여러 사례 중 fountain 부호와 RS 부호를 사용한 연구 결과가 있다 [2]. 이는 DNA 저장장치에 오류정정부호를 효율적으로 적용하면 적은 수의 시퀀싱 리드 수로부터 데이터 완전 복원이 가능함을 보여준 결과이다. [2]의 부호화 방식을 활용하여 더욱 더 복호화 방식을 개선한 연구도 진행되었다 [3]. 두 연구 모두 강판정 복호화 방식을 기반으로 진행되었는데 본 논문에서는 FASTQ 파일과 통계 정보를 활용한 log likelihood ratio (LLR) 계산 방식을 제안하고 이로부터 연판정 복호화를 수행하여 복호화 성능 개선을 시퀀싱 리드 수 관점으로 보인다.

Ⅱ. 본론

본 논문에서는 [3]의 부호화 방식으로부터 얻어진 시퀀싱 결과를 활용하여 연판정 기반의 새로운 복호화 알고리즘을 제안한다. 즉, 동일한 부호화 방식에서 더 효율적인 복호화 알고리즘을 제안한다. 구체적인 부호화 방식과 시퀀싱 환경은 [3]에서 확인할 수 있으며 총 18000 개의 DNA 시퀀스들이 저장되어 있고 각 길이는 primer 를 제외하고 152 이다. 각 DNA 시퀀스들은 각자의 내부에서 RS 부호로 부호화 되어 있다. 더불어 18000 개의 DNA 시퀀스들은 (18000, 16050) LT 부호로 부호화 되어 있는 product 부호 구조이다. [3]에서는 homopolymer runlength 및 GC-content 를 고려하여 설계한 DNA pool 과 그렇지 않은 pool 이 존재하는데 본 논문에서는 seed 정보를 안다는 가정을 하기 위해 제약조건을 고려하지 않은 DNA pool 을 사용하였다.

시퀀싱 결과로 얻을 수 있는 FASTQ 파일은 각 염기가 무엇인지를 나타내는 염기 콜과 해당 염기가 맞을 확률을 제공한다. 그런데 염기는 4 종류이기 때문에 이진 비트 관점에서 한 염기는 2 개의 비트에 대응된다. 따라서 비트 관점의 LLR 계산을 위해서는 A, C, G, T 4 가지 확률이 모두 필요한데 FASTQ 파일에서는 1 가지의 확률만 제공하므로 본 논문에서는 전체 시퀀싱 결과와 저장한 DNA 시퀀스를 편집 거리로 대응시켜 각 DNA 시퀀스 포지션 별로 조건부 오류 확률 얻었다. 예시로 염기 콜이 A 일 때 각 염기의 조건부 오류 확률은

<그림 1>과 같다. 이 때 [3]의 부호화 방식이 삽입 오류나 삭제 오류를 대응하지 못하는 방식이므로 PEAR 알고리즘을 거친 후 길이가 정확히 152 로 일치하는 DNA 시퀀스들만 복호화에 사용하는데 조건부 오류 확률을 위한 통계도 마찬가지로 PEAR 알고리즘 이후 길이가 정확히 152 로 일치하는 DNA 시퀀스들만 활용하였다.



<그림 1. 염기 콜이 A 일 때 DNA 시퀀스 포지션 별 조건부 오류 확률>

위와 같이 염기 콜이 b 일 때 실제로 저장한 염기가 b' 일 조건부 확률을 $p(b'|\text{base call is } b)$ 이라고 표현한다. 이로부터 임의의 염기 b 에 대해 확률 $p(b)$ 가 제공되면 나머지 염기 b' 의 확률은 <그림 1>과 같은 포지션 별 조건부 오류 확률을 사용하여 다음의 수식 $(1-p(b)) \times p(b'|\text{base call is } b)$ 를 통해 얻을 수 있다. A, C, G, T 가 각각 비트 00, 01, 10, 11 에 대응되도록 부호화를 진행하였기 때문에 이에 맞게 각 비트 별로 LLR 을 계산할 수 있다. 가령 첫번째 비트의 경우 LLR 은 $\log\frac{p(\text{A})+p(\text{C})}{p(\text{G})+p(\text{T})}$ 로 얻어진다. 이러한 LLR 계산을 바탕으로 저장한 seed 와 일치하며 그 값이 같은 DNA 리드들을 묶어 최종 LLR 을 계산할 수 있다. 더불어 DNA 시퀀스에는 RS 패리티도 존재하는데 RS 복호화의 경우 강판정을 기반으로 진행하므로 이 경우는 단순히 seed 가 같은 DNA 리드들에 대해 $p(b)$ 의 확률들을 곱하여 제일 큰 염기로 판정하였다. 본 논문에서 최종적으로 제안하는 연판정 기반 복호화 알고리즘은 <알고리즘 1>과 같다.

입력: PEAR 알고리즘 이후 올바른 seed 정보를 갖는 DNA 시퀀스들에 대한 LLR 값, RS 패리티, 재복호화 허용 횟수 n_{re}.
출력: 16050×256 정보 비트.
초기화: $i = 0$

1 단계: 256 개의 이진 LT 부호의 연판정 복호화 수행.
2 단계: 복호화된 각 DNA 시퀀스들에 대해 강판정 기반 RS 복호화 수행.
3 단계: 모든 DNA 시퀀스들이 seed 부분에 오류가 없이 RS 복호화를 성공할 경우 정보 비트들을 복원하고 종료.
4 단계: $i = n_{\text{re}}$일 경우 복호화 실패를 선언하고 종료.
5 단계: $i < n_{\text{re}}$이고 RS 복호화가 seed 부분에 오류가 있다고 주장하거나 RS 복호화 자체가 실패한 경우 해당 DNA 시퀀스를 제거하고 초기 LLR 값을 이용하여 1 단계를 다시 진행. 그 후 i를 1 증가 시킴.

<알고리즘 1. 제안하는 연판정 기반의 복호화 방식>

이로부터 총 DNA 시퀀스 리드 중 특정 수 만큼을 50 번 임의 추출하여 복호화 성공 횟수를 측정한 결과는 <표 1>과 같다.

임의 추출 리드 수	강판정 방식 [3]	연판정 방식
72000	11	17
74000	22	39
76000	38	45
78000	42	48
80000	45	50
82000	49	50
84000	49	50
86000	50	50

<표 1. 강판정 [3] 및 연판정 기반의 복호화 방식 성능 비교>

완전 복원이 성공한 지점의 임의 추출 리드 수 뿐만 아니라 모든 임의 추출 리드 수에 대해 제안하는 연판정 기반의 복호화 방식을 이용할 경우 더 많이 복호화에 성공한 것을 확인할 수 있다. 이는 DNA 저장장치를 활용하는 데에 있어서 제안하는 연판정 기반의 복호화 방식이 기존 강판정 기반의 복호화 방식 대비 저비용, 고신뢰성을 갖는다는 것을 의미한다.

Ⅲ. 결론

본 논문에서는 FASTQ 파일과 통계 정보를 활용한 LLR 계산 방식을 제안하여 연판정 복호화를 수행하였다. 이로부터 기존 강판정 기반의 복호화 방식보다 뛰어난 복호화 성능을 보이는 것을 확인하였다.

ACKNOWLEDGMENT

본 연구는 한국연구재단을 통해 미래창조과학부의 미래유망 융합기술 파이오니어사업으로부터 지원받아 수행되었습니다 (2022M3C1A3090859).

참 고 문 헌

[1] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis, "PEAR: a fast and accurate Illumina Paired-End reAd mergeR," *Bioinformatics*, vol. 30, no. 5, pp. 614-620, 2014.

[2] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950-954, 2017.

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