파운틴 코드를 이용한 DNA 저장 장치에서의 효율적인 복호화를 위한 기법

정재호, 노종선, 박호성*

서울대학교 전기정보공학부 뉴미디어통신공동연구소

*전남대학교 컴퓨터정보통신공학과

jaehoj@ccl.snu.ac.kr, jsno@snu.ac.kr, *hpark1@jnu.ac.kr

Efficient Decoding Method for DNA Storage System using Fountain Codes

Jaeho Jeong, Jong-Seon No, Hosung Park*

INMC, Department of Electrical and Computer Engineering, Seoul National University

*Department of Computer Engineering, Chonnam National University

요 약

본 논문은 파운틴 코드를 이용한 DNA 저장 장치에서 사용될 수 있는 효율적인 복호화 기법에 대해서 알아본다. 먼저, 우리의 이전 연구[2]에서 사용된 DNA 저장 장치의 구조 및 오류정정부호의 스펙에 대해 알아보고, 여기서 새롭게 추가된 Hamming-distance 기반의 기법들과 복호화 순서 책정 방법 등을 자세히 소개한다.

Ⅰ. 서 론

DNA 저장 장치[1]는 차세대 데이터 센터에 사용될만한 새로운 저장 매체로 연구되고 있다. 기존의 디지털 데이터에서 2 bit 당 하나의 염기로 대응(ex: A=00, C=01, G=10, T=11)시켜 데이터를 DNA 형태로 변환하여 저장하는 방식인 DNA 저장 장치는 단 몇 그램에 수백 페타바이트를 저장할 수 있을 정도로 DNA 합성 기술이 발전하고 있고, 기존에 데이터 센터에서 사용되고 있는 자기테이프, 하드디스크, SSD(Solid State Drive) 등에 비해서 같은 부피 대비 백만 배 이상의 높은 용량 효율성을 보이고 있다[5]. 하지만 아직까지는 DNA 데이터를 합성, 보관, 시퀀싱을 하는 과정에서 생물학적 혹은 화학적 요인으로 오류가 필연적으로 발생하게 되어 오류정정부호 등을 이용하여 데이터를 부호화 및 복호화하는 과정이 필수적으로 포함되어야 한다[1]. 이와 관련하여 본 저자진은 최근 파운틴 코드(Fountain code)를 이용하여 데이터를 DNA로 부호화하여 저장하고 이를 다시 복호화해내는 데에 성공하였고[2], 기존에 있던 연구[3]보다 더 적은 양의 올리고(oligo) 샘플들로 원래 파일을 복원해내었기에 DNA 저장 장치의 복호화 성능을 발전시키는 데에 기여하였다. 이 과정에서 오류정정부호를 조심스럽게 설계하고 이와 밀접하게 연계되는 여러 가지 기법들이 필요하였다.

본 논문에서는 파운틴 코드를 이용한 DNA 저장 장치에서 사용될 수 있는 효율적인 복호화 기법에 대해서 알아본다. 먼저 우리의 이전 연구[2]에서 사용된 DNA 저장 장치의 구조 및 오류정정부호의 스펙에 대해 알아보고, 여기서 새롭게 사용된 Hamming-distance 기반 그룹화 방법(clustering)이나 오류 탐지 및 제거 방법(discarding), 새로운 복호화 순서 책정 방법(decoding RS corrected later) 등을 자세히 소개한다.

Ⅱ. 본론

연구 [2]에서 사용된 DNA 저장 장치의 구조에 대해서 알아보면 513.6KB의 이미지 파일을 256bit * 16050 조각으로 분할하여 DNA 형태로 저장하였다. DNA 저장 장치에서의 일반적인 Inner / Outer Code 구조[1]를 사용하였고, inner code로는 RS(Reed-Solomon) code, outer code로는 파운틴 코드 중에서 LT(Luby transform) code를 사용하였다. 이 같은 구조는 기존에 있던 연구[3]에서 실제로 합성 및 시퀀싱 실험을 성공한 바 있었기에 같은 구조를 사용하면서 세부 변수들만 바꿔서 실험을 진행하였다. 2 bit를 1 nt(nucleotide)로 대응시켜 각 올리고는 payload 128nt, seed 16nt, RS code 8nt로 총 길이 152nt로 이루어졌으며, 전체 16050개의 시퀀스를 LT code를 이용하여 18000종류로 구현하였다. 이를 그림으로 나타내면 다음과 같다.

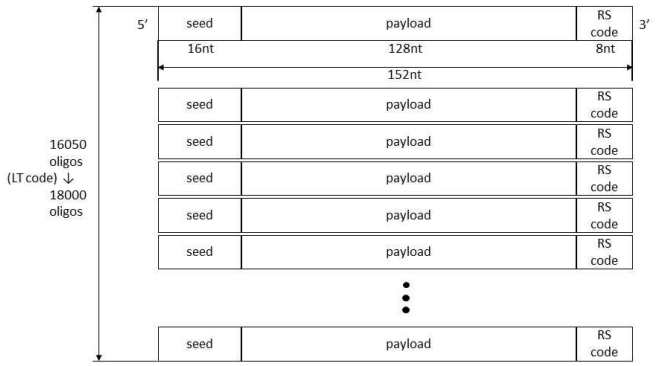

그림1. DNA 저장 장치에 사용된 DNA 올리고 구조

위와 같은 구조를 통해 데이터 파일을 부호화하여 DNA 형태로 저장하는 데 성공하였고, 이를 읽어내는 데에는 Illumina 업체의 시퀀싱 장비를

사용하였다. 이 과정에서 저장된 데이터를 효율적으로 복호화해내기 위해 여러 가지 기법이 필요하였는데, 그중 몇 가지를 살펴보도록 하자.

A. Hamming-distance 기반 그룹화 방법 (Hamming-distance based clustering)

DNA 저장 장치의 데이터를 읽어내기 위해서는 PCR(polymerase chain reaction) 증폭 과정을 거치게 된다. 이때, 같은 종류의 올리고들이 무작위로 복사되는 과정에서 어떤 시퀀스는 수십 개 정도로 많은 숫자가 생성되기도 하지만 어떤 시퀀스는 한두 개만, 혹은 어떤 시퀀스는 구조에 따라서는 아예 생성이 안 되기도 한다.

기존 연구[3]에서는 올리고 길이 전체의 시퀀스가 모두 동일한 것들끼리만 같은 cluster로 묶은 후, cluster의 크기가 큰 순서대로 모아서 복호화를 진행하였는데, 시퀀스 중간에 오류가 하나라도 포함이 되어 있으면 다른 cluster로 묶이게 되고, 또 오류가 포함된 시퀀스가 PCR 과정에서 많이 복제되었다면 오류가 있는 시퀀스가 더 높은 우선순위의 cluster가 되어 부호화 과정에서 잘못된 정보를 전달할 수가 있었다. 우리는 이같이 잘못된 정보가 전달되는 것을 방지하기 위하여, 기존에는 전체 길이가 모두 동일한 시퀀스들만 같은 cluster로 묶었다면, 우리는 Hamming-distance가 2 정도까지 차이가 나는 시퀀스들까지 같은 cluster로 묶은 뒤, 한 cluster 내에 존재하는 여러 시퀀스들에 대해서는 각 자리별로 다수결의 원칙을 이용하여 대표 시퀀스(consensus)를 정하는 방식을 채택하였다. 실제 실험 과정에서는 오류가 겨우 하나 혹은 두 개만 생성된 시퀀스들이 매우 많이 생성되는데, 이 시퀀스들은 해당 오류 한두 개를 제외하고는 모두 정확한 정보를 가지고 있는 것이기에 이들을 버리지 않고 같은 cluster로 묶으면서 최대한 정확한 정보를 대표 시퀀스를 통해 가져올 수 있도록 설계하였고, 그 결과 복호화 성능을 향상시키는 데에 기여하였다.

B. Hamming-distance 기반 오류 탐지 및 제거 방법 (Hamming-distance based discarding)

이전 연구까지는 데이터를 부호화하여 DNA로 합성할 때 거치는 제약사항으로 homopolymer-run (AAAA 등과 같이 염기가 연속으로 나오는 것) 혹은 GC-content (한 올리고 내에서 GC와 AT의 비율을 거의 비슷하게 맞추는 것) 정도만 고려되었다[4]. 하지만 우리는 각각의 올리고들이 모두 Hamming-distance로 약 80nt 정도의 거리를 두도록 설계하였고, 이를 복호화 과정에서 오류를 판별하는 데에 사용하였다. 실제 실험 진행 과정에서는 매우 오류가 많이 포함된, 복호화 과정에 절대로 사용되어서는 안 되는 시퀀스들이 일부 생성되는데 이들이 복호화 과정에 들어가게 된다면 복호화 성능에 치명적인 영향을 끼치게 되므로 이를 어떻게 걸러내느냐가 매우 중요해지게 된다. 우리는 어떤 특정 시퀀스가 기존에 모아놓은 시퀀스 cluster들과 비교했을 때, 거리가 80이 아닌 40, 50 정도만 차이가 나는 시퀀스가 존재한다면 이는 매우 많은 오류가 포함된 시퀀스라고 판단하여 복호화 과정에서 아예 배제하였다. 실제로 이같은 기법을 사용했을 때 오류가 포함된 시퀀스들을 많이 발견할 수 있었고, 이를 걸러내어 복호화 성능을 높이는 데에 이바지할 수 있었다.

C. 복호화 순서 책정 방법 (decoding RS-corrected sequence later)

위에서 언급한 A, B 기법을 사용하여 clustering 과정이 끝난 후에는 RS code의 복호화 과정을 먼저 거친 이후, LT code의 복호화 과정을 시작하게 된다. 최종적으로 LT code의 복호화 과정에 얼마나 정확한 정보들이 어떤 순서로 더 먼저 들어가는지가 매우 중요하게 되는데, 기존 연구[3]에서는 inner code로 RS code를 사용하면서 RS syndrome check를 했을 때 통과되는 시퀀스들, 혹은 오류정정까지 되는 시퀀스들을 모두 복호화 과정에 사용하였다. 하지만 RS code 스펙에 따라서 오류정정이 가능한 범위보다 더 많은 수의 오류가 발생하게 된다면 이는 잘못된 RS codeword로 정정이 되어 LT 복호화에 들어가서 치명적인 영향을 끼칠 수가 있게 된다. 그래서 우리는 RS check 결과가 올바른 시퀀스들을 우선적으로 LT code 복호화 과정에 포함시키고 그 이후에 RS code 오류정정을 한 시퀀스들이 들어가도록 순서를 조정하였고, 이 또한 복호화 성능을 높이는데 기여할 수 있었다.

Ⅲ. 결론

본 논문에서는 파운틴코드를 이용한 DNA 저장 장치에서 복호화 성능을 높일 수 있는 여러 가지 기법들을 소개하였다. 아직까지는 DNA 저장 장치를 실험하는 비용이 비싸기 때문에 실제 실험을 통한 검증은 파운틴 코드와 RS 코드를 이용한 구조에 대해서만 할 수 있었지만, 파운틴 코드뿐만 아니라 clustering 단계가 필요한 모든 오류정정부호에 대해서 사용이 가능한 기법들이다. 또한, 기존의 복호화 방식에 사용된 컴퓨터 연산을 생각했을 때, 올리고 시퀀스들간의 Hamming-distance 비교만을 이용한 기법들이 사용되었기 때문에 연산량도 많이 추가되지 않은 상황에서 DNA 저장 장치의 복호화 성능을 많이 향상시킬 수 있다는 장점이 있다.

논문 [2]에서 진행한 연구는 우리가 DNA 저장 장치에 대해 진행한 첫 번째 연구였기에 더 많은 실험적인 요소를 추가하지 못하였다. 최근의 DNA 저장 장치 연구에서 고려되는 것처럼 seed 부분에 오류정정부호를 추가한다든지, 혹은 차세대 시퀀싱 업체인 Oxford Nanopore 장비를 이용하거나 새로운 합성 방식을 사용하는 등 연구 분야를 확장시킬 수 있는 범위는 매우 넓다. 우리는 이렇게 데이터 합성, 보관, 시퀀싱을 통해 성공적으로 부호화 및 복호화하는 작업까지 모두 진행해 본 경험을 살려서 앞으로도 DNA 저장 장치 연구 분야에 있어서 더 많은 기여를 하고자 한다.

ACKNOWLEDGMENT

본 연구는 삼성미래기술육성센터의 지원을 받아 수행되었음. (SRFC-IT1802-09).

참 고 문 헌

[1] 정재호(Jaeho Jeong),and 노종선(Jong-Seon No). "DNA 저장 매체에 사용되는 오류정정부호에 관한 고찰." 한국통신학회 학술대회논문집 2020.8 (2020): 1113-1114.

[2] Jeong, Jaeho, et al. "Cooperative Sequence Clustering and Decoding for DNA Storage System with Fountain Codes." Bioinformatics (2020).

[3] Erlich, Yaniv, and Dina Zielinski. "DNA Fountain enables a robust and efficient storage architecture." Science 355.6328 (2017): 950-954.

[4] Ross, Michael G., et al. "Characterizing and measuring bias in sequence data." Genome biology 14.5 (2013): 1-20.

[5] Dong, Yiming, et al. "DNA storage: research landscape and future prospects." National Science Review 7.6 (2020): 1092-1107.