DNA 저장 매체에 사용되는 오류정정부호에 관한 고찰

정재호, 노종선
서울대학교 전기정보공학부 뉴미디어통신공동연구소
jaehoj@ccl.snu.ac.kr, jsno@snu.ac.kr

Study about Error Correction Code applied in DNA Storage

Jaeho Jeong and Jong-Seon No
INMC, Department of ECE, Seoul National University

요 약

본 논문은 차세대 저장 매체로 떠오르고 있는 DNA 저장 매체의 구조에 대해 알아보고, DNA 저장 매체에 오류정정부호가 적용되는 방식에 대해 알아본다. 그리고 현재 DNA 저장 매체 기술들이 연구하고 있는 오류정정부호의 종류들을 소개하며 이 부호들이 어떤 논문에서 어떻게 사용되었는지를 정리하였다.

Ⅰ. 서 론

DNA(Deoxyribonucleic Acid, 디옥시리보 핵산) 저장 매체란 데이터를 DNA 형식으로 저장하고, 그렇게 합성한 DNA 염기들 자체를 보관하는 기술을 말한다. 컴퓨터가 0과 1 두 개의 비트(bit)로 데이터를 저장한다면 DNA 저장 매체는 A(Adenine, 아데닌), C(Cytosine, 사이토신), G(Guanine, 구아닌), T(Thymine, 티민) 총 4종류의 염기를 이용하여 데이터를 저장하는데, 일반적으로는 컴퓨터 데이터를 바이너리(binary) 데이터로 나타낸 뒤, 2비트 당 1개의 염기로 대응시켜 DNA를 합성한다고 보면 된다(ex: A = 00, C = 01, G = 10, T = 11). 데이터를 DNA로 저장하는 합성(synthesis) 과정, 보관된 DNA를 복제시키는 PCR 증폭 과정, 저장된 DNA의 데이터를 읽어내는 시퀀싱(sequencing) 과정 등에서 생물학적 요인들에 의해 필연적으로 오류가 발생하는데, 이를 해결하기 위해 DNA에 오류정정부호를 적용하여 합성하는 방안이 연구되고 있다. 그러나 아직까지는 DNA 저장 매체 기술이 기초 수준에 머물러 있기 때문에 여러 가지 다양한 오류정정부호들을 적용하여 실험한 논문들이 발표되고 있고, 각각의 실험 결과들을 통해 어떤 부호들이 어떤 상황에서 활용되면 좋은지, 그 부호들의 구조나 성능은 어떤지 등이 연구되고 있다.

본 논문에서는 DNA 저장 매체의 전체적인 구조를 먼저 알아보고, DNA 저장 매체의 채널(channel) 특성 및 현재까지 DNA 저장 매체와 관련하여 발표된 논문들에서 사용한 다양한 오류정정부호들이 어떻게 적용되었는지를 간단히 살펴볼 예정이다.

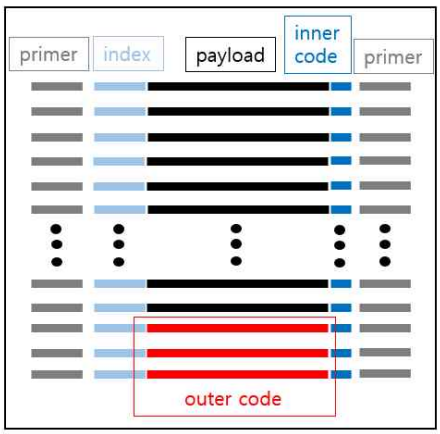
Ⅱ. 본 론

A.DNA저장매체의구조

현재까지 DNA를 인위적으로 합성시키는 기술에는 시간이 오래 걸리고 비용이 많이 필요하며 길이가 길어질수록 오류율이 급격하게 증가한다는 단점이 존재한다. 그래서 아직까지는 한 DNA 안에 염기의 개수가 100~300개 정도가 되도록 DNA를 합성하고 있으며, 이렇게 합성된 DNA 한 가닥을 올리고(oligo)라고 부른다.

일반적으로 길이 100짜리 올리고 하나에는 총 200개의 비트 정보가 저장될 수 있기 때문에, 컴퓨터로 100KByte 정도 되는 데이터를 저장하기 위해서는 단순계산으로도 약 4000종류의 길이 100 올리고로 데이터를 쪼개서 합성해야 하고 이를 payload라 부른다. 하지만 이 올리고들을 기존 데이터로 제대로 복원을 하기 위해서는, DNA 형식으로 저장하고 읽어내는 전체 과정에서 필연적으로 발생하는 오류를 방지하기 위해 오류정정부호를 적용하여 패리티(parity) 정보를 같이 포함시켜야 된다. 그래서 수천 종류의 전체 올리고들을 모았을 때 이 올리고들이 제대로 모아졌는지를 확인하기 위한 outer code, 길이 100~300 정도의 각각의 올리고 내부의 오류들을 찾아내기 위한 inner code가 필요하게 된다.

실제로는 저장하려는 데이터의 크기에 따라 수천~수십만 종류의 올리고를 한 번에 합성하게 된다. 여기서 각각의 올리고들을 순서대로 재배열해야 우리가 처음 저장할 때 의도했던 데이터를 복원할 수 있기 때문에 각각의 올리고에는 그 순서를 저장하는 인덱스(index)가 추가되어야 한다. 이후, 최종적으로는 데이터 구분을 위해 필요한 프라이머(primer), 혹은 어댑터(adapter)가 올리고 양 끝에 붙어서 합성되게 된다. 이렇게 DNA 저장 매체의 전체 구조를 도식도로 나타낸 것이 <그림 1>이다.



<그림 1> DNA 저장 매체 구조

B. 오류정정부호의 적용

현재 DNA 저장 매체 분야에서 수많은 오류정정부호가 적용되고 있지만, 가장 보편적으로 사용되는 부호들은 몇 개 정해져 있다. 어떤 상황에서 어떤 부호들이 주로 쓰이는지, 또 각각의 부호는 어떤 특징을 가지고 있는지 간단히 알아보겠다.

Inner Code:

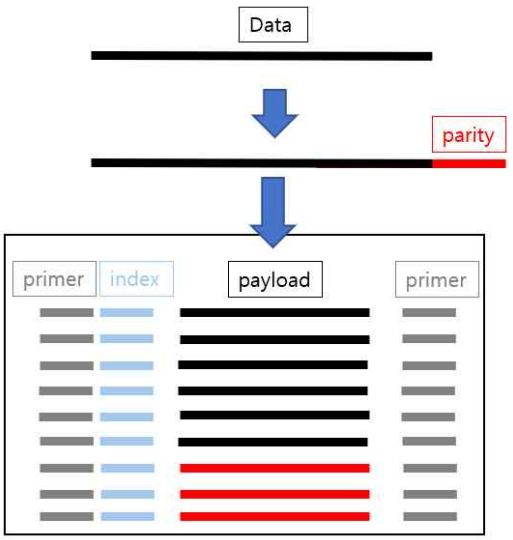
Inner Code는 코드 길이가 100~300 정도로 짧고, 통상적인 통신시스템에서 발생하는 대체 오류(Substitution Error)뿐만 아니라 삽입 오류(Insertion Error), 삭제 오류(Deletion Error)도 발생한다는 특징이 있다. 짧은 길이에서 효율적으로 적용 가능한 부호들이 사용되는데 최근에는 RS(Reed-Solomon) Code가 주로 사용되고 있다[1][4]. 현재까지 대체 오류를 잡아내는 데에는 큰 어려움이 없으나, 삽입 오류와 삭제 오류가 동시에 발생하면서 이를 긴 길이의 대체 오류로 착각하게 되는 문제를 해결하기 위해 많은 연구가 진행되고 있다.

Outer Code:

Outer Code는 코드 길이가 수천~수십만 정도로 길고, 인덱스의 개수가 정해져 있기 때문에 대체 오류만 존재한다. 또한, 중간 과정에서 몇몇 인덱스의 올리고들이 통째로 손실되는 경우가 발생하기도 한다. 길이가 긴 코드에 적합하거나, 혹은 디코딩에 실패한 경우 시퀀싱한 데이터를 조금만 더 가져와서 추가 계산만 하면 되는 부호면 좋기 때문에 반복적인(iterative) 디코딩 계산이 가능한 부호가 연구되고 있다. 그래서 Fountain Code의 일종인 LT Code가 사용되거나[1], 혹은 Inner Code와 마찬가지로 RS Code가 사용되기도 한다[2][4].

Single Large Block Code:

[3]번 논문에서는 특이하게도 오류정정부호를 inner code 방향과 outer code 방향 두 개로 나눠서 사용한 것이 아닌, 전체 데이터를 하나로 묶어서 하나의 긴 길이의 코드로 사용하였다. 대신에, 오류가 연속적으로 발생해도 정정이 가능하고 매우 긴 길이에서 성능이 좋은 부호 중의 하나인 LDPC(Low Density Parity Check) Code가 사용되었다. 일반적인 inner/outer code 구조가 아닌 하나의 large block code 구조를 그림으로 나타내면 아래 <그림 2>의 구조와 같다.



<그림 2> Single Large Block Code 구조

Ⅲ. 결 론

위에 나와 있는 부호들 이외에도 다른 부호들이 여러 가지 용도로 사용되기도 한다. 간단한 수준의 DNA 합성만이 진행되었기에 해밍(Hamming)부호가 사용되었거나[5], 인덱스 부분에만 거는 용도로 BCH 부호가 사용되는 등[3] 아직까지 수많은 오류정정부호가 DNA 저장매체에 접목되어 실험되고 있고, 각각의 부호들에 대해 이를 복호해 내기 위한 여러 기법들이 소개되고 있다. 또한, 위에 언급된 오류정정부호들에 있어서도 각 부호에 필요한 코드 비율(code rate)이나 변수(parameter)값을 어떻게 설정하느냐에 따라서도 성능이 매우 달라지고, 혹은 오류정정에 필요한 계산을 경판정 복호(hard decoding) 방식으로 적용하느냐[1] 아니면 연판정 복호(soft decoding) 방식으로 적용하느냐[3]에 따라서도 오류정정부호 적용 기법이 완전히 달라질 수 있다.

이외에도 DNA 저장 매체를 읽어내는 시퀀싱 장비를 Illumina Sequencing 방식을 사용하는 것과 Oxford Nanopore Sequencing 방식을 사용하는 것에 따라서도 DNA 저장 매체의 채널 특성이 완전히 다르기 때문에 각각의 환경에 맞는 오류정정부호를 찾기 위한 노력도 많이 들어가고 있다. 이 논문을 통해 DNA 저장 매체의 근본적인 구조와 적용되는 오류정정부호의 종류 들을 알게 되면 차후 DNA 저장 매체에 관한 연구와 더불어 DNA 저장 매체에 적용되는 새로운 기법들을 이해하는 데 큰 도움이 될 것이다.

ACKNOWLEDGMENT

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

참 고 문 헌

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

[2] Organick, Lee, et al. "Random access in large-scale DNA data storage." Nature biotechnology 36.3 (2018): 242.

[3] Chandak, Shubham, et al. "Improved read/write cost tradeoff in DNA-based data storage using LDPC codes." 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2019.

[4] Meiser, Linda C., et al. "Reading and writing digital data in DNA." Nature Protocols 15.1 (2020): 86-101.

[5] Takahashi, Christopher N., et al. "Demonstration of end-to-end automation of DNA data storage." Scientific reports 9.1 (2019): 1-5.