(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0048362
(43) 공개일자 2022년04월19일

(51) 국제특허분류(Int. Cl.)
H03M 7/14 (2006.01) *G06N 3/12* (2006.01)
H03M 13/37 (2006.01) *H03M 7/28* (2006.01)
(52) CPC특허분류
H03M 7/14 (2013.01)
G06N 3/123 (2013.01)
(21) 출원번호 10-2020-0131431
(22) 출원일자 2020년10월12일
심사청구일자 2020년10월12일

(71) 출원인
서울대학교산학협력단
서울특별시 관악구 관악로 1 (신림동)
전남대학교산학협력단
광주광역시 북구 용봉로 77 (용봉동)
(뒷면에 계속)
(72) 발명자
정재호
경기도 성남시 분당구 양현로 272, 310동 1301호
박호성
광주광역시 광산구 수완로33번길 76, 106-101
(뒷면에 계속)
(74) 대리인
특허법인비엘티

전체 청구항 수 : 총 15 항

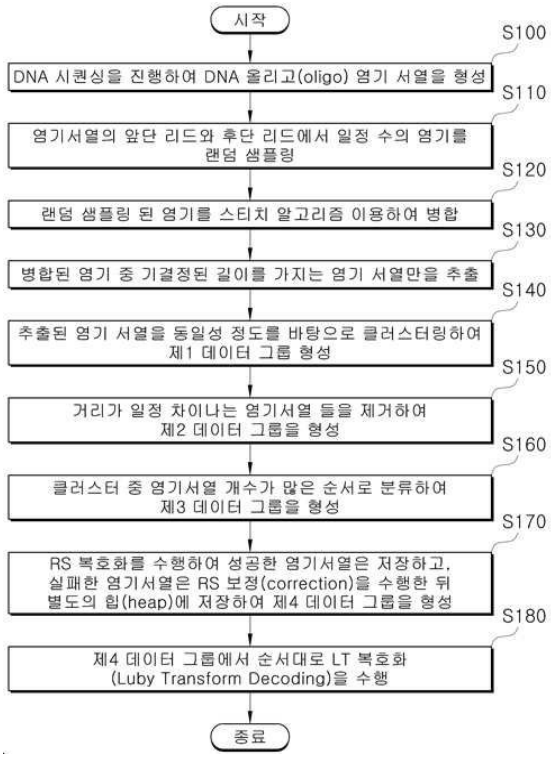
(54) 발명의 명칭 **DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법, 프로그램 및 장치**

(57) 요 약

본 발명은 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법에 관한 것으로, DNA 시퀀싱을 진행하여 DNA 올리고(oligo) 염기 서열을 형성하는 단계, 염기 서열의 앞단 리드(forward read)와 후단 리드(reverse read)에서 각각 일정 수의 염기를 랜덤하게 샘플링 하는 단계, 랜덤 샘플링된 염기를 스티치 알고리즘(stitch algorithm)을

(뒷면에 계속)

대 표 도 - 도1



이용하여 합치는 단계, 스티치 알고리즘을 통해 합쳐진 염기 서열들 중 기결정된 길이를 가지는 염기 서열만을 추출하는 단계, 상기 추출된 염기 서열을 동일성 정도를 바탕으로 클러스터링하여 제1 데이터 그룹을 형성하는 단계, 상기 제1 데이터 그룹에서 클러스터들과 거리가 일정 차이나는 염기서열 들을 제거하여 제2 데이터 그룹을 형성하는 단계, 상기 제2 데이터 그룹에서 클러스터 중 염기서열 개수가 많은 순서로 분류하여 제3 데이터 그룹을 형성하는 단계, 상기 제3 데이터 그룹에서 염기서열 개수가 많은 클러스터부터 RS 복호화를 수행하여 성공한 염기서열은 저장하고, 실패한 염기서열은 RS 보정(correction)을 수행한 뒤 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성하는 단계, 및 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 단계;를 포함한다.

"본 발명은 삼성전자 미래기술육성센터의 지원을 받아 수행된 연구임 (과제번호 SRFC-IT1802-09)"

(52) CPC특허분류
H03M 13/3761 (2013.01)
H03M 7/28 (2013.01)
(71) 출원인
울산대학교 산학협력단
울산광역시 남구 대학로 93(무거동)
홍익대학교 산학협력단
서울특별시 마포구 와우산로 94 (상수동)

(72) 발명자
박성준
서울특별시 서초구 잠원동 신반포로 33길 66, 101-606
노종선
서울특별시 강남구 광평로10길 15 상록수아파트 109동 406호
김성환
서울특별시 광진구 아차산로66길 70, 102-302
노 알버트
서울시 용산구 한강대로 205 용산파크자이아파트 D동 2317호

명 세 서

청구범위

청구항 1

전자장치가 DNA 시퀀싱을 진행하여 DNA 올리고(oligo) 염기 서열을 형성하는 단계;

상기 전자장치가 생성된 염기 서열의 앞단 리드(forward read)와 후단 리드(reverse read)에서 각각 일정 수의 염기를 랜덤하게 샘플링 하는 단계;

상기 전자장치가 랜덤 샘플링된 염기를 스티치 알고리즘(stitch algorithm)을 이용하여 합치는 단계;

상기 전자장치가 스티치 알고리즘을 통해 합쳐진 염기 서열들 중 기결정된 길이를 가지는 염기 서열만을 추출하는 단계;

상기 전자장치가 상기 추출된 염기 서열을 동일성 정도를 바탕으로 클러스터링하여 과반수 이상으로 모인 DNA 염기서열 베이스를 바탕으로 대표값을 형성하고, 과반수 미만의 DNA 염기서열 베이스는 단순 저장하여 제1 데이터 그룹을 형성하는 단계;

상기 전자장치가 상기 제1 데이터 그룹에서 클러스터들과 거리가 일정 차이나는 염기서열 들을 제거하여 제2 데이터 그룹을 형성하는 단계;

상기 전자장치가 상기 제2 데이터 그룹에서 클러스터 중 염기서열 개수가 많은 순서로 분류하여 제3 데이터 그룹을 형성하는 단계;

상기 전자장치가 상기 제3 데이터 그룹에서 염기서열 개수가 많은 클러스터부터 RS 복호화를 수행하여 성공한 염기서열은 저장하고, 실패한 염기서열은 RS 보정(correction)을 수행한 뒤 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성하는 단계; 및

상기 전자장치가 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 단계; 를 포함하는 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법.

청구항 2

제 1 항에 있어서,

상기 DNA 올리고 염기 서열을 형성하는 단계는,

서로 해밍 거리(Hamming distance)가 최소 길이 x 이상으로 차이나도록 형성되는 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법.

청구항 3

제 2 항에 있어서,

상기 제2 데이터 그룹을 형성하는 단계는,

상기 제1 데이터 그룹에서 클러스터들과 거리가 3 nt 내지 최소 길이 x-2 nt 만큼 차이나는 염기서열 들을 제거하는 단계인, DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법.

청구항 4

제 1 항에 있어서,

상기 제1 데이터 그룹을 형성하는 단계는,

기결정된 길이를 가지는 염기 서열 중 기결정된 길이 전체가 동일한 염기 서열, 1개 염기 서열이 상이하고 나머지 서열은 동일한 염기 서열 및 2개 염기 서열이 상이하고 나머지 서열은 동일 한 염기 서열까지 동일성이 있는 것으로 판단하여 동일 클러스터로 결정하는 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법.

청구항 5

제 1 항에 있어서,

상기 제3 데이터 그룹을 형성하는 단계는,

클러스터 중에서 개수가 1개인 것들의 순서를 결정하는 방법은, 예측품질점수(Q-score)를 바탕으로 연산한 염기를 잘못 불러올 확률 값 P가 높은 순서로 결정하는 것인 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법.

청구항 6

제 5 항에 있어서,

상기 확률 값 P는 하기의 수학식을 만족하는 값인 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법.

[수학식 1]

$$P = \prod_{k=1}^{y}(1 - 10^{-\frac{Q(k)}{10}})$$

여기에서, y는 염기 서열을 추출하는 기결정된 염기 서열의 길이를 의미한다.

청구항 7

제 1 항에 있어서,

상기 전자장치가 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 단계는,

RS 복호화에 성공한 염기서열에 대하여 먼저 LT 복호화를 수행하고, 이후에 RS 보정을 수행한뒤 heap에 저장된 염기서열에 대하여 LT 복호화를 수행하는 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법.

청구항 8

컴퓨터인 하드웨어와 결합되어, 제1항 내지 제7항 중 어느 한 항의 방법을 실행하기 위해 매체에 저장된, DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 프로그램.

청구항 9

DNA 시퀀싱을 진행하여 DNA 올리고(oligo) 염기 서열을 형성하고,

생성된 염기 서열의 앞단 리드(forward read)와 후단 리드(reverse read)에서 각각 일정 수의 염기를 랜덤하게 샘플링 하고,

랜덤 샘플링된 염기를 스티치 알고리즘(stitch algorithm)을 이용하여 합치고,

스티치 알고리즘을 통해 합쳐진 염기 서열들 중 기결정된 길이를 가지는 염기 서열만을 추출하고,

상기 추출된 염기 서열을 동일성 정도를 바탕으로 클러스터링하여 과반수 이상으로 모인 DNA 염기서열 베이스를 바탕으로 대표값을 형성하고, 과반수 미만의 DNA 염기서열 베이스는 단순 저장하여 제1 데이터 그룹을 형성하고,

상기 전자장치가 상기 제1 데이터 그룹에서 클러스터들과 거리가 일정 차이나는 염기서열 들을 제거하여 제2 데이터 그룹을 형성하고,

상기 전자장치가 상기 제2 데이터 그룹에서 클러스터 중 염기서열 개수가 많은 순서로 분류하여 제3 데이터 그룹을 형성하고,

상기 전자장치가 상기 제3 데이터 그룹에서 염기서열 개수가 많은 클러스터부터 RS 복호화를 수행하여 성공한 염기서열은 저장하고, 실패한 염기서열은 RS 보정(correction)을 수행한 뒤 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성하고,

상기 전자장치가 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 DNA 저장 장치

의 시퀀스 집단화 방식 기반 복호화 컴퓨팅 장치.

청구항 10

제 9 항에 있어서,

상기 DNA 올리고 염기 서열을 형성하는 단계는,

서로 해밍 거리(Hamming distance)가 최소 길이 x 이상으로 차이나도록 형성되는 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 컴퓨팅 장치.

청구항 11

제 10 항에 있어서,

상기 제2 데이터 그룹을 형성하는 단계는,

상기 제1 데이터 그룹에서 클러스터들과 거리가 3 nt 내지 최소 길이 x-2 nt 만큼 차이나는 염기서열 들을 제거하는 단계인, DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 컴퓨팅 장치.

청구항 12

제 9 항에 있어서,

상기 제1 데이터 그룹을 형성하는 단계는,

기결정된 길이를 가지는 염기 서열 중 기결정된 길이 전체가 동일한 염기 서열, 1개 염기 서열이 상이하고 나머지 서열은 동일한 염기 서열 및 2개 염기 서열이 상이하고 나머지 서열은 동일 한 염기 서열까지 동일성이 있는 것으로 판단하여 동일 클러스터로 결정하는 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 컴퓨팅 장치.

청구항 13

제 9 항에 있어서,

상기 제3 데이터 그룹을 형성하는 단계는,

클러스터 중에서 개수가 1개인 것들의 순서를 결정하는 방법은, 예측품질점수(Q-score)를 바탕으로 연산한 염기를 잘못 불러올 확률 값 P가 높은 순서로 결정하는 것인 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 컴퓨팅 장치.

청구항 14

제 13 항에 있어서,

상기 확률 값 P는 하기의 수학식을 만족하는 값인 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 컴퓨팅 장치.

[수학식 1]

$$P = \prod_{k=1}^{y} \left(1 - 10^{-\frac{Q(k)}{10}}\right)$$

여기에서, y는 염기 서열을 추출하는 기결정된 염기 서열의 길이를 의미한다.

청구항 15

제 9 항에 있어서,

상기 전자장치가 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 단계는,

RS 복호화에 성공한 염기서열에 대하여 먼저 LT 복호화를 수행하고, 이후에 RS 보정을 수행한뒤 heap에 저장된 염기서열에 대하여 LT 복호화를 수행하는 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 컴퓨팅 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법, 프로그램 및 장치에 관한 것이다.

배 경 기 술

[0002] 현대 사회에서는 스마트 기기들의 보급, SNS 사용량의 증가 및 IoT 시스템의 보급 등으로 데이터가 넘쳐나고 있으며, 2020년이면 44조 GB 이상의 데이터가 인터넷에 존재할 것으로 예상된다.

[0003] 이와 같이 많은 양의 데이터는 일반적으로 하드웨어 장치 자체, 외장 하드 및 웹 하드 등 많은 저장 장치를 이용하여 곳곳에 데이터를 보관하고 있다.

[0004] 그러나, 이러한 데이터의 보관 방식은 용량이나 비용뿐만 아니라 불편함까지 초래하는 많은 한계들이 있어, 최근 DNA 스토리지 기술에 대한 많은 연구가 이루어지고 있다.

[0005] DNA 스토리지 기술이란, DNA의 염기를 0과 1로 치환하여 디지털 데이터를 저장하는 방식으로서, DNA에 유전이 아닌 디지털 데이터를 저장한다면 1g의 DNA에 4550억 GB 정도의 데이터를 저장할 수 있게 되며, 이는 1TB 하드 디스크보다 215,000배 더 많은 데이터 저장 능력인 것이다.

[0006] 따라서 최근, 각 디지털 데이터를 DNA 데이터로 인코딩 하는 방법과 DNA 데이터를 디지털 데이터로 복호화 하는 방법에 대한 연구가 계속해서 이루어지고 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국 공개특허공보 제 10-2012-0031838호

발명의 내용

해결하려는 과제

[0008] DNA 시퀀스 집단화 방식에 있어서, Erlich, Y. 및 Zielinski의 선행 연구에 따르면 스티치 알고리즘을 통해 합성된 염기 서열이 완벽히 동일한 경우만 동일 클러스터로 판단하도록 하고 있다. 이러한 방식은 개선의 여지가 있는 매우 적은 오류를 가지는 클러스터들조차 복호화에 활용하지 못함으로써, 복호화의 효율 이 저하되고 오히려 오류값이 높아지는 문제점이 발생하고 있다.

[0009] 상술한 바와 같은 문제점을 해결하기 위한 본 발명은 DNA 합성에 있어서,

[0010] 스티치 알고리즘을 통해 형성된 염기 서열에서 특정 길이를 가지는 염기 서열을 추출할 때 동일성 정도를 바탕으로 클러스터링을 진행하되, 완전히 동일한 염기 서열 뿐 아니라 적은 에러를 가지는 염기 서열들도 클러스터에 포함시킴으로써 에러 데이터를 활용하여 복호화 효율을 향상시키고, 최종 에러율을 감소시키는 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법, 프로그램 및 장치를 제공하고자 한다.

[0011] 또한, 본 발명은 RS 복호화 수행시 성공한 클러스터 뿐 아니라 실패한 클러스터도 RS 보정(correction)을 진행하여 LT 복호화에 활용함으로써 종래에는 버려졌던 정보를 복구하여 사용함으로써 복호화의 효율성 향상이 가능한, DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법, 프로그램 및 장치를 제공하고자 한다.

[0012] 또한, 본 발명은 예측품질점수(Q-score) 값을 바탕으로 LT 복호화의 우선순위를 결정함으로써 복호화의 에러율 저하가 가능한, DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법, 프로그램 및 장치를 제공하고자 한다.

[0013] 본 발명이 해결하고자 하는 과제들은 이상에서 언급된 과제로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0014] 상술한 과제를 해결하기 위한 본 발명의 일 실시예에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법은, 전자장치가 DNA 시퀀싱을 진행하여 DNA 올리고(oligo) 염기 서열을 형성하는 단계, 상기 전자장치가 생성된 염기 서열의 앞단 리드(forward read)와 후단 리드(reverse read)에서 각각 일정 수의 염기를 랜덤하게 샘플링 하는 단계, 상기 전자장치가 랜덤 샘플링된 염기를 스티치 알고리즘(stitch algorithm)을 이용하여 합치는 단계, 상기 전자장치가 스티치 알고리즘을 통해 합쳐진 염기 서열들 중 기결정된 길이를 가지는 염기 서열만을 추출하는 단계, 상기 전자장치가 상기 추출된 염기 서열을 동일성 정도를 바탕으로 클러스터링하여 과반수 이상으로 모인 DNA 염기서열 베이스를 바탕으로 대표값을 형성하고, 과반수 미만의 DNA 염기서열 베이스는 단순 저장하여 제1 데이터 그룹을 형성하는 단계, 상기 전자장치가 상기 제1 데이터 그룹에서 클러스터들과 거리가 일정 차이나는 염기서열 들을 제거하여 제2 데이터 그룹을 형성하는 단계, 상기 전자장치가 상기 제2 데이터 그룹에서 클러스터 중 염기서열 개수가 많은 순서로 분류하여 제3 데이터 그룹을 형성하는 단계, 상기 전자장치가 상기 제3 데이터 그룹에서 염기서열 개수가 많은 클러스터부터 RS 복호화를 수행하여 성공한 염기서열은 저장하고, 실패한 염기서열은 RS 보정(correction)을 수행한 뒤 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성하는 단계 및 상기 전자장치가 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 단계를 포함한다.

[0015] 상기 DNA 올리고 염기 서열을 형성하는 단계는,

[0016] 서로 해밍 거리(Hamming Distance)가 최소 길이 x 이상으로 차이나도록 형성되는 것이다.

[0017] 상기 제2 데이터 그룹을 형성하는 단계는, 상기 제1 데이터 그룹에서 클러스터들과 거리가 3 nt 내지 최소 길이 x-2 nt 만큼 차이나는 염기서열 들을 제거하는 것이다.

[0018] 상기 제1 데이터 그룹을 형성하는 단계는, 기결정된 길이를 가지는 염기 서열 중 기결정된 길이 전체가 동일한 염기 서열, 1개 염기 서열이 상이하고 나머지 서열은 동일한 염기 서열 및 2개 염기 서열이 상이하고 나머지 서열은 동일한 염기 서열까지 동일성이 있는 것으로 판단하여 동일 클러스터로 결정하는 것이다.

[0019] 상기 제3 데이터 그룹을 형성하는 단계는, 클러스터 중에서 개수가 1개인 것들의 순서를 결정하는 방법은, 예측 품질점수(Q-score)를 바탕으로 연산한 염기를 잘못 불러올 확률 값 P가 높은 순서로 결정하는 것이다.

[0020] 상기 확률 값 P는 하기의 수학식을 만족하는 값인 것이다.

[0021] **[수학식 1]**

[0022] $$P = \prod_{k=1}^{y}\left(1 - 10^{-\frac{Q(k)}{10}}\right)$$

[0023] 여기에서, y는 염기 서열을 추출하는 기결정된 염기 서열의 길이를 의미한다.

[0024] 상기 전자장치가 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 단계는, RS 복호화에 성공한 염기서열에 대하여 먼저 LT 복호화를 수행하고, 이후에 RS 보정을 수행한뒤 heap에 저장된 염기서열에 대하여 LT 복호화를 수행하는 것이다.

[0025] 상술한 과제를 해결하기 위한 본 발명의 다른 실시예에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 프로그램은 컴퓨터인 하드웨어와 결합되어, 상기 방법 중 어느 하나의 방법을 실행하기 위해 매체에 저장된다.

[0026] 상술한 과제를 해결하기 위한 본 발명의 다른 실시예에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 컴퓨팅 장치는,

[0027] DNA 시퀀싱을 진행하여 DNA 올리고(oligo) 염기 서열을 형성하고, 생성된 염기 서열의 앞단 리드(forward read)와 후단 리드(reverse read)에서 각각 일정 수의 염기를 랜덤하게 샘플링 하고, 랜덤 샘플링된 염기를 스티치 알고리즘(stitch algorithm)을 이용하여 합치고, 스티치 알고리즘을 통해 합쳐진 염기 서열들 중 기결정된 길이를 가지는 염기 서열만을 추출하고, 상기 추출된 염기 서열을 동일성 정도를 바탕으로 클러스터링하여 과반수 이상으로 모인 DNA 염기서열 베이스를 바탕으로 대표값을 형성하고, 과반수 미만의 DNA 염기서열 베이스는 단순 저장하여 제1 데이터 그룹을 형성하고, 상기 전자장치가 상기 제1 데이터 그룹에서 클러스터들과 거리가 일정 차이나는 염기서열 들을 제거하여 제2 데이터 그룹을 형성하고, 상기 전자장치가 상기 제2 데이터 그룹에서 클러스터 중 염기서열 개수가 많은 순서로 분류하여 제3 데이터 그룹을 형성하고, 상기 전자장치가 상기 제3 데이터 그룹에서 염기서열 개수가 많은 클러스터부터 RS 복호화를 수행하여 성공한 염기서열은 저장하고, 실패한 염

기서열은 RS 보정(correction)을 수행한 뒤 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성하고, 상기 전자장치가 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행한다.

[0028] 상기 DNA 올리고 염기 서열을 형성하는 것은,

[0029] 서로 해밍 거리가 최소 길이 x 이상으로 차이나도록 형성되는 것이다.

[0030] 상기 제2 데이터 그룹을 형성하는 것은, 상기 제1 데이터 그룹에서 클러스터들과 거리가 3 nt 내지 최소 길이 x-2 nt 만큼 차이나는 염기서열 들을 제거하는 것이다.

[0031] 상기 제1 데이터 그룹을 형성하는 것은, 기결정된 길이를 가지는 염기 서열 중 기결정된 길이 전체가 동일한 염기 서열, 1개 염기 서열이 상이하고 나머지 서열은 동일한 염기 서열 및 2개 염기 서열이 상이하고 나머지 서열은 동일한 염기 서열까지 동일성이 있는 것으로 판단하여 동일 클러스터로 결정하는 것이다.

[0032] 상기 제3 데이터 그룹을 형성하는 것은, 클러스터 중에서 개수가 1개인 것들의 순서를 결정하는 방법은, 예측품질점수(Q-score)를 바탕으로 연산한 염기를 잘못 불러올 확률 값 P가 높은 순서로 결정하는 것이다.

[0033] 상기 확률 값 P는 하기의 수학식을 만족하는 값인 것이다.

[0034] **[수학식 1]**

[0035] $$P = \prod_{k=1}^{y}(1 - 10^{-\frac{Q(k)}{10}})$$

[0036] 여기에서, y는 염기 서열을 추출하는 기결정된 염기 서열의 길이를 의미한다.

[0037] 상기 전자장치가 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 것은, RS 복호화에 성공한 염기서열에 대하여 먼저 LT 복호화를 수행하고, 이후에 RS 보정을 수행한뒤 heap에 저장된 염기서열에 대하여 LT 복호화를 수행하는 것이다.

[0038] 이 외에도, 본 발명을 구현하기 위한 다른 방법, 다른 시스템 및 상기 방법을 실행하기 위한 컴퓨터 프로그램을 기록하는 컴퓨터 판독 가능한 기록 매체가 더 제공될 수 있다.

발명의 효과

[0039] 상기와 같은 본 발명에 따르면,

[0040] 스티치 알고리즘을 통해 형성된 염기 서열에서 특정 길이를 가지는 염기 서열을 추출할 때 동일성 정도를 바탕으로 클러스터링을 진행하되, 완전히 동일한 염기 서열 뿐 아니라 적은 에러를 가지는 염기 서열들도 클러스터에 포함함으로써 에러 데이터를 활용하여 복호화 효율을 향상시키고, 최종 에러율을 감소시킬 수 있다.

[0041] 본 발명은 RS 복호화 수행시 성공한 클러스터 뿐 아니라 실패한 클러스터도 RS 보정(correction)을 진행하여 LT 복호화에 활용함으로써 종래에는 버려졌던 정보를 복구하여 사용함으로써 복호화의 효율성 향상이 가능하다.

[0042] 본 발 명은 복호화 에러율의 큰 원인 중 하나인 크기 1짜리 클러스터에 대하여 예측품질점수(Q-score) 값을 바탕으로 LT 복호화의 우선순위를 결정함으로써 복호화의 에러율 저하가 가능하다.

[0043] 본 발명의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0044] 도 1은 본 발명의 일 실시예에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법을 설명하기 위한 흐름도이다.

도 2a 및 2b는 본 발명의 일 실시예에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법과 일 비교예에 따른 복호화 방법의 성능을 비교한 도면이다.

도 3은 Q-score의 확률 값이 낮을수록 에러율이 높아지는 상관관계를 도시한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0045] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야의 통상의 기술자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0046] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다. 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 발명의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0047] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0048] 설명에 앞서 본 명세서에서 사용하는 용어의 의미를 간략히 설명한다. 그렇지만 용어의 설명은 본 명세서의 이해를 돕기 위한 것이므로, 명시적으로 본 발명을 한정하는 사항으로 기재하지 않은 경우에 본 발명의 기술적 사상을 한정하는 의미로 사용하는 것이 아님을 주의해야 한다.

[0049] 본 명세서에서 '해독 염기 서열 파일'은 DNA 염기 서열을 나타내는 텍스트 기반의 표준 염기 데이터 형식의 파일이며, 예컨대, FASTQ 파일을 의미한다.

[0050] 본 발명은 시퀀스 집단화 방식을 통해 DNA 데이터로 인코딩된 데이터를 다시 복호화 하는 과정에 대한 것이다.

[0051] 따라서, 본 명세서에서는 인코딩된 데이터를 복호화 하는 과정에 대하여 기술하고 있으며, 다양한 방법으로 인코딩된 데이터를 모두 포함할 수 있다.

[0052] 특히, 본 발명은 루비 트랜스폼(LT, Luby Transform) 코드, 파운틴(Fountain) 코드, 터보(Turbo) 코드, 폴라(Polar) 코드 또는 LDPC(Low Density Parity Check) 코드로 인코딩된 데이터를 복호화하는 것에 적합할 수 있으나, 특정 방식으로 인하여 인코딩된 데이터를 북호화하는 것으로 한정하지는 않고, 어떠한 방식으로든 인코딩된 데이터에 대하여 모두 적용될 수 있다.

[0053] 본 발명인 복호화 방법의 설명을 돕기 위하여, 특정 데이터를 DNA 데이터로 인코딩하는 방법에 대하여 간략하게 설명한다.

[0054] 하기의 예시는, 루비 트랜스폼 코드를 이용하여 인코딩하는 일반적인 방법의 예시이다.

[0055] 특정 이미지 파일을 DNA 데이터로 인코딩 하는 과정에서, 하나의 이미지를 길이 L의 K_{LT}개의 패킷으로 나눈다. 나뉘어진 K_{LT}개의 패킷을 루비 트랜스폼 인코딩을 통하여 하나의 인코딩된 패킷으로 생성한다. 이후, 선형 피드백 시프트 레지스터(LFSR, Linear-Feedback Shift Register)에 의해 얻어진 시드 비트를 붙이고, 특정 값 이상에 해당하는 RS(Reed Solomon) 인코딩 비트를 붙인다.

[0056] 최종 인코딩된 비트에 대하여 각각 염기의 비트로 변환한 후, 염기 서열의 최대 호모 폴리머 길이가 미리 정해진 길이보다 작거나 같고, 구아닌, 시토신 함량 비가 미리 정해진 비율에 속하는지를 확인한다.

[0057] 상기 조건이 만족된다면, 해당 서열은 인코딩되도록 선택되고, 상기 조건이 만족되지 않는 서열의 경우에는 폐기된다. 상기와 같은 절차를 모든 데이터에 대하여 서열이 합성될때까지 반복함으로써, DNA 데이터로 인코딩되는 것이다.

[0058] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세하게 설명한다.

[0059] 도 1은 본 발명의 일 실시예에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법을 설명하기 위한 흐름도이다.

[0060] 도 1을 참조하면, 본 발명의 일 실시예에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법은 DNA 시퀀싱을 진행하여 DNA 올리고(oligo) 염기 서열을 형성하는 단계(S100), 염기서열의 앞단 리드와 후단 리드에서 일정 수의 염기를 랜덤 샘플링하는 단계(S110), 랜덤 샘플링 된 염기를 스티치 알고리즘 이용하여 병합하는 단계(S120), 병합된 염기 중 기결정된 길이를 가지는 염기 서열만을 추출하는 단계(S130), 추출된 염기 서열을 동일성 정도를 바탕으로 클러스터링하여 제1 데이터 그룹 형성하는 단계(S140), 거리가 일정 차이나는 염기서열들을 제거하여 제2 데이터 그룹을 형성하는 단계(S150), 클러스터 중 염기서열 개수가 많은 순서로 분류하여 제3 데이터 그룹을 형성하는 단계(S160), RS 복호화를 수행하여 성공한 염기서열은 저장하고, 실패한 염기서열은 RS 보정(correction)을 수행한 뒤 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성하는 단계(S170), 및 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 단계(S180)을 포함한다.

[0061] DNA 시퀀싱을 진행하여 DNA 올리고(oligo) 염기 서열을 형성하는 단계(S100)에서, 전자장치는 시퀀스 식별자로서 각각의 염기인, 아데닌(A, Adenine), 구아닌(G, Guanine), 시토신(C, Cytosine) 및 티민(T, Thymine)을 나타내는 정보와 시퀀스의 순서를 포함하는 강성 정보를 바탕으로 DNA 올리고 염기 서열을 형성한다.

[0062] 랜덤 샘플링 된 염기를 스티치 알고리즘 이용하여 병합하는 단계(S120)는, 전자 장치가 염기 서열 파일에 대하여 스티치 알고리즘을 이용하여 기결정된 타겟 길이의 스티치된 시퀀스를 생성하는 것이다.

[0063] 스티치 알고리즘(PEAR)은 쌍방향 리드의 겹침을 평가하는 알고리즘으로서, 병렬 처리되는 것이 특징이다.

[0064] 타겟 길이는 스티치 알고리즘에서 허용되는 길이일 수 있으며, 기결정된 길이, 예를 들어, 152 nt의 길이일 수 있다.

[0065] 병합된 염기 중 기결정된 길이를 가지는 염기 서열만을 추출하는 단계(S130)는, 상기 스티치된 시퀀스에서 해밍 거리(Hamming distance)를 기준으로 인코딩된 시퀀스를 기결정된 길이, 예를 들어 152 nt, 를 가지도록 추출하는 것이다.

[0066] 상기 해밍 거리는 미리 정해진 최소 편집 거리를 기준으로 하는 것이며, 해당 편집 거리를 기준으로 하여 스티치된 올리고 리드의 초기 시퀀스를 추출한다. 상기 해밍 거리는 예를 들어, 80 nt일 수 있다.

[0067] 추출된 염기 서열을 동일성 정도를 바탕으로 클러스터링하여 제1 데이터 그룹 형성하는 단계(S140)는, 기결정된 길이를 y라고 할 때 염기 서열 중 y - 2 개까지 일치하는 시퀀스를 동일한 클러스터로 집단화한다. 예를 들어, 기결정된 길이를 152 nt일 경우, 염기 서열 중 150 nt까지 동일한 시퀀스, 즉 2개까지 다른 시퀀스는 동일한 클러스터로 모을 수 있다. 기존의 기술에서는 기결정된 염기서열 모두가 완벽히 동일한 경우에 클러스터로 모으고, 단 1개의 염기서열이라도 다른 경우는 클러스터에 포함시키지 않았으나, 본 발명에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법은 2개까지의 차이는 클러스터로 포함시키는 것을 특징으로 한다. 따라서, 한 클러스터에는 시퀀스가 완벽히 동일하거나, 1개만 다르거나, 2개까지 다른 것이 공존할 수 있다. 에러가 1, 2개 정도 있는 DNA 올리고를 동일 클러스터에 모음으로써 해당 DNA 올리고의 에러를 위치별 다수결 원칙으로 확인할 수 있다. 에러 2개까지 차이나는 것들을 동일 클러스터 에 포함시킨 이유는 다수결의 원칙으로 에러를 찾아내지 못하는 위치의 경우 RS code를 이용하여 해당 에러를 확인할 수 있기 때문이다. 에러를 확인하는 방법은 후술하도록 한다.

[0068] 추출된 모든 시퀀스에 대하여 클러스터링이 완료된 경우에는, 각 클러스터마다 염기 위치별로 과반수 이상으로 모인 DNA 염기가 있을 경우 해당 염기로 통일시킨 대표 값을 형성한다. 만약 과반수에 해당하는 염기가 없는 위치의 경우에는 해당 위치에 가능한 모든 염기값을 남긴다.

[0069] 본 단계(S140)에서 클러스터링된 시퀀스 데이터를 제1 데이터 그룹으로 형성한다.

[0070] 거리가 일정 차이나는 염기서열 들을 제거하여 제2 데이터 그룹을 형성하는 단계(S150)는, 제1 데이터 그룹을 기준으로 클러스터들과 거리가 일정 차이나는 염기서열 들을 제거한다.

[0071] 예를 들어, 상기 일정거리는 최소 해밍 거리를 x라 할 때, 3 내지 x-2 만큼을 포함할 수 있다. 즉, 클러스터들과 거리가 3 내지 x-2만큼 차이나는 DNA 올리고는 제1 데이터 그룹에서 제거한다. 예를 들어, 편집 거리가 80 nt 일 경우, 제거되는 DNA 올리고는 인접 클러스터와 3 nt ~ 78 nt만큼 거리차이가 나는 DNA 올리고이다.

[0072] 이런 염기서열을 제거하는 이유는 DNA 올리고 합성 시 최소 간격인 편집 거리를 설정했음에도, 그 이하의 해밍 거리를 가지는 DNA 올리고는 매우 많은 오류를 포함할 수 있는 가능성이 높기 때문이다. 다만, 1nt, 2nt, 그리고 x-1, x-2 의 차이나는 DNA 올리고는 삭제하지 않고 남기는 것은 RS 보정을 통해 정정할 수 있기 때문이다.

[0073] 클러스터 중 염기서열 개수가 많은 순서로 분류하여 제3 데이터 그룹을 형성하는 단계(S160)는, 제2 데이터 그룹을 기준으로 클러스터 중에서 개수가 많은 것에서 적은 순서로 배열(sorting)하여 제3 데이터 그룹을 형성한다.

[0074] 클러스터 개수가 1인 염기 서열 들은 복호화 단계에서 에러에 많은 영향을 주므로, 해당 염기 서열에 대한 LT 복호화의 우선순위를 결정하기 위해 예측품질점수(Q-score)를 바탕으로 연산한 염기를 잘못 불러올 확률 값 P가 높은 순서로 결정하여 제3 데이터 그룹을 형성할 수 있다.

[0075] 확률 값 P는 하기의 수학식을 만족하는 값일 수 있다.

[0076] **[수학식 1]**

[0077] $$P = \prod_{k=1}^{y}(1 - 10^{-\frac{Q(k)}{10}})$$

[0078] 여기에서, y는 염기 서열을 추출하는 기결정된 염기 서열의 길이를 의미한다.

[0079] 예를 들어, 기결정된 염기 서열의 길이가 152 nt인 경우, $P = \prod_{k=1}^{152}(1 - 10^{-\frac{Q(k)}{10}})$ 일 수 있다. 이 수학식은 염기 서열의 각 위치가 해당 염기로 시퀀싱 되었을 때, 신뢰할 수 있는지에 대한 확률을 나타내는 수치이다. 이를 전체, 염기 서열에 대해 곱함으로써 해당 DNA 올리고 전체에 대한 신뢰성을 나타내는 지표로써 P값은 이용될 수 있다.

[0080] RS 복호화를 수행하여 성공한 염기서열은 저장하고, 실패한 염기서열은 RS 보정(correction)을 수행한 뒤 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성하는 단계(S170)는,

[0081] 종래 기술에는 RS 복호화 결과가 실패한 클러스터들은 전혀 사용하지 않았다. 그러나 본 발명에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법은, 오류가 겨우 1개 포함된 cluster들의 경우에는 RS 보정을 통해 올바른 값으로 수정이 가능하고 LT 복호화에 활용하도록 하였다. 따라서, 본 발명은 기존의 버려진 정보를 복구하여 LT 복호화에 사용할 수 있도록 하였고, RS 복호화 결과가 옳은 클러스터들을 우선으로 실패한 애들은 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성하도록 하였다.

[0082] 예를 들어 oligo X와 Y가 2개만큼 차이나는데 동일 클러스터로 집단화 된 경우, 일차는 다수결의 원칙을 통해 올바른 올리고인지를 판단한다. 만일 동률을 이뤄서 X, Y 중 어느것이 맞는 oligo인지 판단할 수 없는 경우에는 RS code를 통해 판별할 수 있다.

[0083] **[Case 1]** X: RS code 통과, Y: RS code 통과

[0084] **[Case 2]** X: RS code 통과, Y: RS code 실패

[0085] **[Case 3]** X: RS code 실패, Y: RS code 실패

[0086] 첫 번째 케이스는, X, Y 모두 RS code를 통과했으므로 X, Y가 일치하는 올리고이므로 둘 중 어떤 시퀀스 든 LT 복호화에 사용할 수 있다.

[0087] 두 번째 케이스는, RS code를 통과한 X가 에러가 없는 온전한 oligo이므로 LT 복호화에 이용한다.

[0088] 세 번째 케이스는, X, Y 둘다 RS를 실패한 경우에는, RS 보정을 진행하여 같은 코드로 보정되는 경우에는 해당 코드를 옳은 코드로 인식하여 힙에 저장한 뒤 낮은 우선순위로 LT 복호화에 이용한다. 만일, 같은 코드로 보정되지 않는 경우에는 둘 중 하나의 시퀀스에 치명적 오류가 있다는 의미이므로 두 시퀀스를 모두 폐기한다.

[0089] 마지막으로 제4 데이터 그룹에서 순서대로 LT 복호화(Luby Transform Decoding)을 수행하는 단계(S180)는,

[0090] 신뢰성이 높은 염기서열부터 낮은 염기서열의 순서로 LT 복호화를 진행함으로써 효율적인 복호화가 가능하다.

[0091] 도 2는 본 발명의 일 실시예에 따른 DNA 저장 장치의 시퀀스 집단화 방식 기반 복호화 방법과 일 비교예에 따른 복호화 방법의 성능을 비교한 도면이다.

[0092] 도 2a 및 2b를 참조하면, 본 발명에 따른 시퀀스 집단화 방식 기반 복호화 방법이 비교예 대비 모든 랜덤 샘플링 회수 구간에서 성능이 우위에 있음을 확인할 수 있다.

[0093] [표 1]

Total 200	Constrained pool		Non-constrained pool	
Random sampling number	Erlich's method	Proposed method	Erlich's method	Proposed method
72000	1	98		
74000	17	147		
76000	89	179	6	138
78000	161	192	56	173
80000	176	196	143	193
82000	190	199	179	200 (*)
84000	196	199	187	198
86000	198	200 (*)	196	200
88000	198	200	198	200
90000	195	200	200 (*)	200
92000	200 (*)	200		

[0094]

[0095] 본 발명의 경우, 제한된 풀에서는 200개의 실험에 대해서 랜덤 샘플링 개수 78000에서는 성공 회수가 190을 도과하고 있으며, 86000부터는 200개 실험 전체 복호화가 성공하는 것을 확인할 수 있다.

[0096] 반면, 비교예의 경우, 제한된 풀에서는 200개의 실험에 대해서 랜덤 샘플링 개수 82000에서야 성공 회수가 190을 도과하고 있으며, 92000부터 200개 실험 전체 복호화가 성공하는 등 본 발명 대비 더 많은 랜덤 샘플링 개수를 요구하는 것을 확인할 수 있다.

[0097] 본 발명의 경우, 제한되지 않은 풀에서는 200개의 실험에 대해서 랜덤 샘플링 개수 80000에서는 성공 회수가 190을 도과하고 있으며, 86000부터는 200개 실험 전체 복호화가 성공하는 것을 확인할 수 있다.

[0098] 반면, 비교예의 경우, 제한되지 않은 풀에서는 200개의 실험에 대해서 랜덤 샘플링 개수 86000에서야 성공 회수가 190을 도과하고 있으며, 90000부터 200개 실험 전체 복호화가 성공하는 등 본 발명 대비 더 많은 랜덤 샘플링 개수를 요구하는 것을 확인할 수 있다.

[0099] [표 2]

		86000 constrained	82000 non-constrained
	Number of sequence reads	73764.7	68744.6
Erlich's method	Number of all clusters	24682.9	24579.9
	Number of size-1 clusters	9443.2	9966.9
	Ratio of size-1 clusters to all clusters	38.3%	40.5%
	Ratio of size-1 clusters to all sequence reads	12.8%	14.5%
Proposed method	Number of all clusters	17634.7	17328.9
	Number of size-1 clusters	2054.2	2431.4
	Ratio of size-1 clusters to all clusters	11.6%	14.0%
	Ratio of size-1 clusters to all sequence reads	2.8%	3.5%
	No. of discarded reads	218.0	401.9
	No. of RS-corrected reads	190.8	222.1

[0100]

[0101] 200개의 실험에 대해서 200개가 모두 성공하는 시점의 세부 데이터를 본 발명과 비교예를 대비할 경우, 개수 1에 해당하는 클러스터의 비율이 비교예 38.3 ~ 40.5% 에서 본 발명 11.6 ~ 14.0% 로 감소함을 확인할 수 있다. 또한, 개수 1에 해당하는 클러스터의 전체 시퀀스 리드에 대한 비율이 12.8% ~ 14.5% 에서 2.8 3.5% 로 감소함을 확인할 수 있다.

[0102] 전술한 바와 같이, 개수 1에 해당하는 클러스터의 비율은 복호화시 에러율이 많은 영향을 미치므로, 이러한 개수 1에 해당하는 클러스터의 비율의 감소는 본 발명에 따른 복호화의 성능상 우위를 입증하는 데이터에 해당한다.

[0103] 또한, RS 보정을 통해 힙에 저장하여 LT 복호화에 재활용한 시퀀스의 개수가 비교예의 경우 0%인 반면, 본 발명의 경우 0.26 ~ 0.32% 정도에 해당하여 정보 활용의 효율성을 증진시킴을 확인할 수 있다.

[0104] 도 3은 Q-score의 확률 값이 낮을수록 에러율이 높아지는 상관관계를 도시한 그래프이다.

[0105] 도 3을 참조하면, $P=\prod_{k=1}^{152}(1-10^{-\frac{Q(k)}{10}})$ 과 해당 하는 시퀀스 리드들이 기존 DNA 올리고 대비 포함하고 있는 오류와의 상관관계가 도시된다. x축은 해당 올리고 리드의 Q-score에 해당하는 확률이며, y축은 올리고 리드와 암호화된 올리고 시퀀스와의 최소 편집 거리 차이를 나타낸다. P값이 높을수록 차이가 적으므로 에러가 적다는 것을 확인할 수 있고, P값이 낮아 질수록 에러가 전반적으로 상승하는 것을 확인할 수 있다.

[0106] 이상에서 전술한 본 발명의 일 실시예에 따른 방법은, 하드웨어인 서버와 결합되어 실행되기 위해 프로그램(또는 어플리케이션)으로 구현되어 매체에 저장될 수 있다.

[0107] 상기 전술한 프로그램은, 상기 컴퓨터가 프로그램을 읽어 들여 프로그램으로 구현된 상기 방법들을 실행시키기 위하여, 상기 컴퓨터의 프로세서(CPU)가 상기 컴퓨터의 장치 인터페이스를 통해 읽힐 수 있는 C, C++, JAVA, 기계어 등의 컴퓨터 언어로 코드화된 코드(Code)를 포함할 수 있다. 이러한 코드는 상기 방법들을 실행하는 필요한 기능들을 정의한 함수 등과 관련된 기능적인 코드(Functional Code)를 포함할 수 있고, 상기 기능들을 상기 컴퓨터의 프로세서가 소정의 절차대로 실행시키는데 필요한 실행 절차 관련 제어 코드를 포함할 수 있다. 또한, 이러한 코드는 상기 기능들을 상기 컴퓨터의 프로세서가 실행시키는데 필요한 추가 정보나 미디어가 상기 컴퓨터의 내부 또는 외부 메모리의 어느 위치(주소 번지)에서 참조되어야 하는지에 대한 메모리 참조관련 코드를 더 포함할 수 있다. 또한, 상기 컴퓨터의 프로세서가 상기 기능들을 실행시키기 위하여 원격(Remote)에 있는 어떠한 다른 컴퓨터나 서버 등과 통신이 필요한 경우, 코드는 상기 컴퓨터의 통신 모듈을 이용하여 원격에 있는 어떠한 다른 컴퓨터나 서버 등과 어떻게 통신해야 하는지, 통신 시 어떠한 정보나 미디어를 송수신해야 하는지 등에 대한 통신 관련 코드를 더 포함할 수 있다.

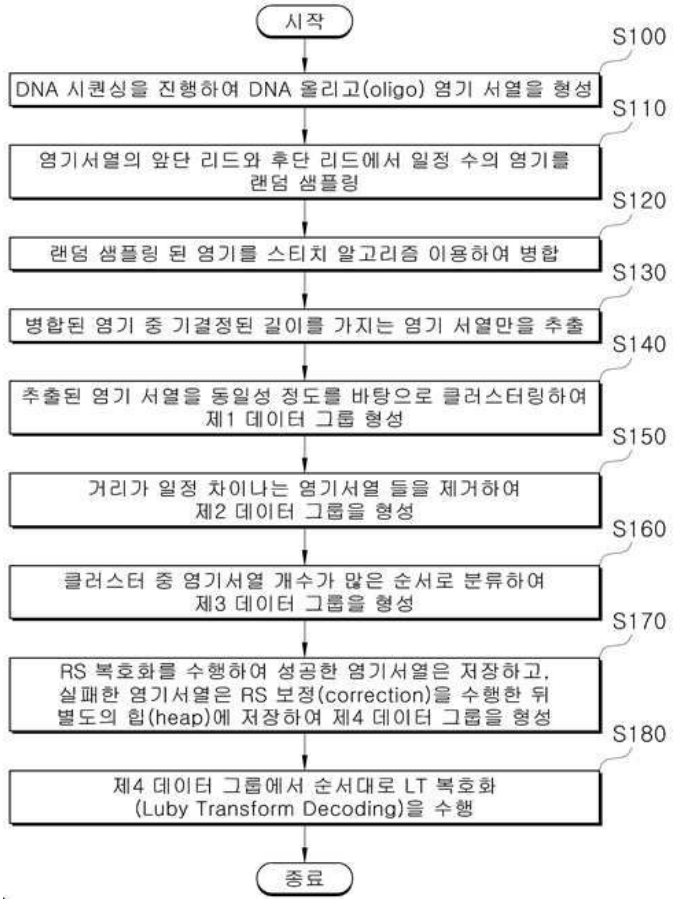
[0108] 상기 저장되는 매체는, 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상기 저장되는 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피디스크, 광 데이터 저장장치 등이 있지만, 이에 제한되지 않는다. 즉, 상기 프로그램은 상기 컴퓨터가 접속할 수 있는 다양한 서버 상의 다양한 기록매체 또는 사용자의 상기 컴퓨터상의 다양한 기록매체에 저장될 수 있다. 또한, 상기 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장될 수 있다.

[0109] 본 발명의 실시예와 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로 직접 구현되거나, 하드웨어에 의해 실행되는 소프트웨어 모듈로 구현되거나, 또는 이들의 결합에 의해 구현될 수 있다. 소프트웨어 모듈은 RAM(Random Access Memory), ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능 기록매체에 상주할 수도 있다.

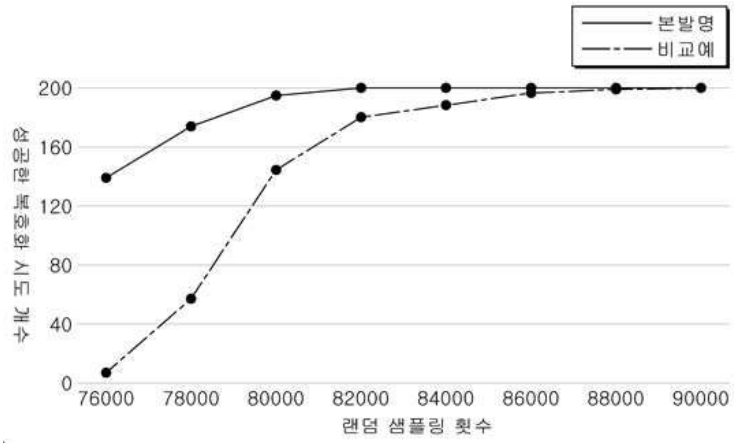
[0110] 이상, 첨부된 도면을 참조로 하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야의 통상의 기술자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이 아닌 것으로 이해해야만 한다.

도면

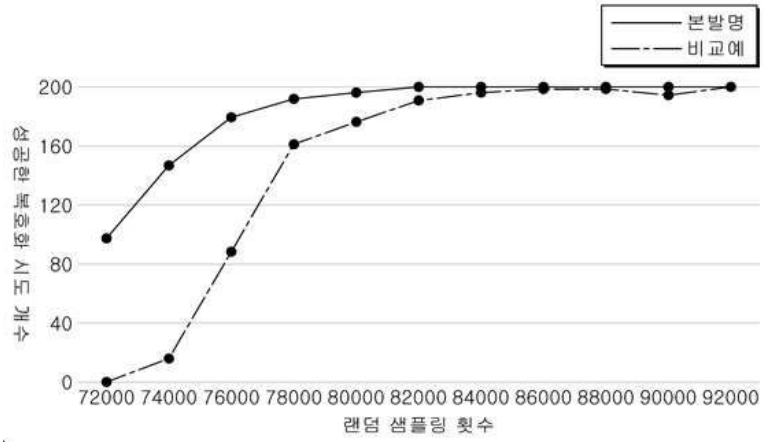
도면1


시작
S100
DNA 시퀀싱을 진행하여 DNA 올리고(oligo) 염기 서열을 형성
S110
염기서열의 앞단 리드와 후단 리드에서 일정 수의 염기를 랜덤 샘플링
S120
랜덤 샘플링 된 염기를 스티치 알고리즘 이용하여 병합
S130
병합된 염기 중 기결정된 길이를 가지는 염기 서열만을 추출
S140
추출된 염기 서열을 동일성 정도를 바탕으로 클러스터링하여 제1 데이터 그룹 형성
S150
거리가 일정 차이나는 염기서열 들을 제거하여 제2 데이터 그룹을 형성
S160
클러스터 중 염기서열 개수가 많은 순서로 분류하여 제3 데이터 그룹을 형성
S170
RS 복호화를 수행하여 성공한 염기서열은 저장하고, 실패한 염기서열은 RS 보정(correction)을 수행한 뒤 별도의 힙(heap)에 저장하여 제4 데이터 그룹을 형성
S180
제4 데이터 그룹에서 순서대로 LT 복호화 (Luby Transform Decoding)을 수행
종료

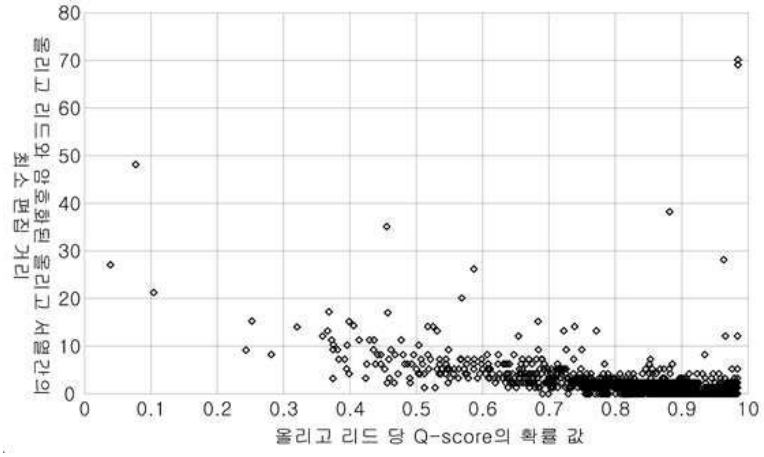
도면2a


본발명
비교예
200
160
120
80
40
0
76000
78000
80000
82000
84000
86000
88000
90000
랜덤 샘플링 횟수

도면2b


본발명
비교예
200
160
120
80
40
0
성공한 복호화 시도 개수
72000 74000 76000 78000 80000 82000 84000 86000 88000 90000 92000
랜덤 샘플링 횟수

도면3


80
70
60
50
40
30
20
10
0
올리고 리드와 복호화된 올리고 리드 사이의 최소 편집 거리
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 10
올리고 리드 당 Q-score의 확률 값