(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0118853
(43) 공개일자 2019년10월21일

(51) 국제특허분류(Int. Cl.)
G16B 30/00 (2019.01)
(52) CPC특허분류
G16B 30/00 (2019.02)
(21) 출원번호 10-2018-0042269
(22) 출원일자 2018년04월11일
심사청구일자 2018년04월11일

(71) 출원인
서울대학교산학협력단
서울특별시 관악구 관악로 1 (신림동)
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
권성훈
서울특별시 강남구 언주로30길 56, C동 1705호(도곡동, 타워팰리스)
박욱
경기도 용인시 수지구 성복2로 184, 203동 1402호(성복동, 성동마을 수지자이2차)
(뒷면에 계속)
(74) 대리인
팬코리아특허법인

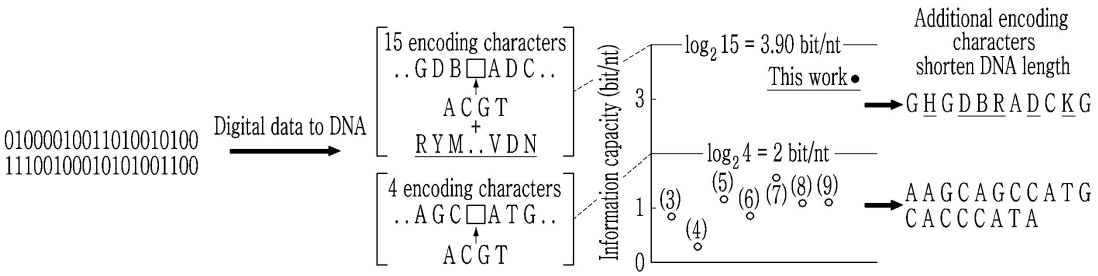
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **DNA 디지털 데이터 저장 장치 및 저장 방법, 그리고 디코딩 방법**

(57) 요 약

DNA 디지털 데이터 저장 방법은 복수의 비트 데이터를 적어도 하나의 디제너레이트 염기를 포함하는 복수의 염기 서열로 인코딩하는 단계, 및 상기 적어도 하나의 디제너레이트 염기를 구성하는 적어도 두 종류의 염기들을 혼합 비율에 기초하여 기판 위에서 합성하는 단계를 포함한다.

대 표 도



(72) 발명자

최영재

서울특별시 서초구 신반포로15길 19, 111동 707호 (반포동, 아크로리버파크)

류태훈

서울특별시 금천구 시흥대로73길 11, 6동 807호(시흥동, 무지개아파트)

송석홍

경기도 용인시 기흥구 용구대로2152번길 7, 104동 303호(신갈동, 고현마을고려에센스빌)

김혜리

경기도 수원시 영통구 청명로59번길 45-22, 301호 (영통동)

김서주

경기도 수원시 영통구 반달로35번길 30, 542호(영통동)

명 세 서

청구범위

청구항 1

복수의 비트 데이터를 적어도 하나의 디제너레이트 염기를 포함하는 복수의 염기 서열로 인코딩하는 단계; 및

상기 적어도 하나의 디제너레이트 염기를 구성하는 적어도 두 종류의 염기들을 혼합 비율에 기초하여 기판 위에서 합성하는 단계를 포함하는 DNA 디지털 데이터 저장 방법.

청구항 2

제1항에 있어서,

상기 복수의 염기 서열 중 단일 종류의 염기를 상기 기판 위에서 합성하는 단계를 더 포함하는 DNA 디지털 데이터 저장 방법.

청구항 3

제1항에 있어서,

상기 적어도 하나의 디제너레이트 염기를 합성하는 단계는,

상기 혼합 비율에 기초한 상기 적어도 두 종류의 염기들을 상기 기판 위에 혼입하는 단계를 포함하는 DNA 디지털 데이터 저장 방법.

청구항 4

제1항에 있어서,

상기 적어도 하나의 디제너레이트 염기를 합성하는 단계는,

외부에서 상기 혼합 비율에 따라 혼합된 상기 적어도 두 종류의 염기들을 상기 기판 위에 혼입하는 단계를 포함하는 DNA 디지털 데이터 저장 방법.

청구항 5

제1항에 있어서,

상기 혼합 비율에서, 상기 적어도 두 종류의 염기들 간의 비율이 동일한 DNA 디지털 데이터 저장 방법.

청구항 6

제1항에 있어서,

상기 혼합 비율에서, 상기 적어도 두 종류의 염기들 간의 비율이 다른 DNA 디지털 데이터 저장 방법.

청구항 7

제1항에 있어서,

상기 적어도 두 종류의 염기들은 DNA 염기, RNA 염기, 및 핵산 유사체(Nucleic acid analogue) 중 적어도 하나인 DNA 디지털 데이터 저장 방법.

청구항 8

복수의 비트 데이터를 인코딩한 복수의 염기 서열이 저장 기본 단위로 구분되어 저장된 DNA 디지털 데이터 저장 장치에 있어서,

상기 복수의 염기 서열 중 상기 저장 기본 단위에 대응하는 제1 염기 서열에 기초하여 복수의 염기들이 합성된 분자들을 포함하고,

상기 제1 염기 서열은 적어도 두 종류의 염기로 구성된 제1 디제너레이트 염기를 포함하며,

상기 분자들은,

상기 적어도 두 종류의 염기 중 제1 염기가 합성된 제1 분자들, 및

상기 적어도 두 종류의 염기 중 제2 염기가 적어도 한 염기 합성된 제2 분자들을 포함하는 DNA 디지털 데이터 저장 장치.

청구항 9

제8항에 있어서,

상기 제1 분자들과 상기 제2 분자들 간의 비율은 혼합 비율에 따르는 DNA 디지털 데이터 저장 장치.

청구항 10

제9항에 있어서,

상기 혼합 비율에서, 상기 적어도 두 종류의 염기들 간의 비율이 동일한 DNA 디지털 데이터 저장 장치.

청구항 11

제9항에 있어서,

상기 혼합 비율에서, 상기 적어도 두 종류의 염기들 간의 비율이 다른 DNA 디지털 데이터 저장 장치.

청구항 12

제8항에 있어서,

상기 적어도 두 종류의 염기들은 DNA 염기, RNA 염기, 및 핵산 유사체(Nucleic acid analogue) 중 적어도 하나인 DNA 디지털 데이터 저장 장치.

청구항 13

복수의 DNA 프래그멘트를 주소에 따라 분류하는 단계;

상기 분류된 복수의 DNA 프래그멘트 각각에 대해 동일한 위치에서의 염기를 분석하는 단계;

상기 분석 결과에 기초하여 염기 비율의 산포도를 산출하는 단계;

상기 염기 비율의 산포도가 적어도 두 염기들의 혼합 비율일 때, 해당 위치의 염기를 디제너레이트 염기로 결정하는 단계; 및

상기 결정된 디제너레이트 염기를기초로 데이터를 디코딩하는 단계를 포함하는 DNA 데이터의 디코딩 방법.

청구항 14

제13항에 있어서,

상기 염기를 분석하는 단계는,

상기 복수의 DNA 프래그멘트 각각의 각 위치에 대한 염기 호출이 분석되는 단계를 포함하는 DNA 데이터의 디코딩 방법.

청구항 15

제13항에 있어서,

상기 염기 비율의 산포도가 적어도 두 염기의 제1 혼합 비율일 때, 해당 위치의 염기를 제1 디제너레이트 염기로 결정하고,

상기 염기 비율의 산포도가 상기 적어도 두 염기의 제2 혼합 비율일 때, 해당 위치의 염기를 제2 디제너레이트 염기로 결정하며,

상기 제1 혼합 비율과 상기 제2 혼합 비율이 서로 다르고, 상기 제1 디제너레이트 염기와 상기 제2 디제너레이트 염기가 서로 다른 DNA 데이터의 디코딩 방법.

발명의 설명

기 술 분 야

[0001] 본 개시는 DNA 디지털 데이터 저장 장치 및 저장 방법, 그리고 DNA 디지털 데이터의 디코딩 방법에 관한 것이다.

배 경 기 술

[0002] 빠른 접근을 위해 모드 데이터를 플래시 메모리에 저장된다고 가정할 때, 2040년에는 디지털데이터 스토리지에 대한 연간 수요가 기하 급수적으로 증가하여 실리콘 공급을 초과할 것으로 예상된다. 디지털데이터의 대량 축적을 위해서, 대체 저장 방법 개발이 필요하다.

[0003] DNA의 높은 물리적 정보 밀도와 내구성으로 인해, DNA를 디지털 데이터 저장 매체로 사용하는 것이 정보 저장에 대한 급격히 증가하는 요구를 해결하는 방법으로 부상했다.

[0004] 그러나 DNA 디지털 데이터 저장 장치가 아직 실제로 구현되고 있지 못하고 있는데, 이는 단위 데이터저장 당 비용이 높기 때문이다.

발명의 내용

해결하려는 과제

[0005] DNA를 이용한 디지털 데이터 저장 비용을 감소시켜 실제적으로 DNA 디지털 데이터 저장 장치 및 저장 방법, 및 디코딩 방법을 구현하고자 한다.

과제의 해결 수단

[0006] 발명의 한 특징에 따른 DNA 디지털 데이터 저장 방법은, 복수의 비트 데이터를 적어도 하나의 디제너레이트 염기를 포함하는 복수의 염기 서열로 인코딩하는 단계, 및 상기 적어도 하나의 디제너레이트 염기를 구성하는 적어도 두 종류의 염기들을 혼합 비율에 기초하여 기판 위에서 합성하는 단계를 포함한다.

[0007] 상기 DNA 디지털 데이터 저장 방법은, 상기 복수의 염기 서열 중 단일 종류의 염기를 상기 기판 위에서 합성하는 단계를 더 포함할 수 있다.

[0008] 상기 적어도 하나의 디제너레이트 염기를 합성하는 단계는, 상기 혼합 비율에 기초한 상기 적어도 두 종류의 염기들을 상기 기판 위에 혼입하는 단계를 포함할 수 있다.

[0009] 상기 적어도 하나의 디제너레이트 염기를 합성하는 단계는, 외부에서 상기 혼합 비율에 따라 혼합된 상기 적어도 두 종류의 염기들을 상기 기판 위에 혼입하는 단계를 포함할 수 있다.

[0010] 상기 혼합 비율에서, 상기 적어도 두 종류의 염기들 간의 비율이 동일할 수 있다.

[0011] 상기 혼합 비율에서, 상기 적어도 두 종류의 염기들 간의 비율이 다를 수 있다.

[0012] 상기 적어도 두 종류의 염기들은 DNA 염기, RNA 염기, 및 핵산 유사체(Nucleic acid analogue) 중 적어도 하나일 수 있다.

[0013] 발명의 다른 특징에 따른 복수의 비트 데이터를 인코딩한 복수의 염기 서열이 저장 기본 단위로 구분되어 저장된 DNA 디지털 데이터 저장 장치는, 상기 복수의 염기 서열 중 상기 저장 기본 단위에 대응하는 제1 염기 서열에 기초하여 복수의 염기들이 합성된 분자들을 포함하고, 상기 제1 염기 서열은 적어도 두 종류의 염기로 구성된 제1 디제너레이트 염기를 포함하며, 상기 분자들은, 상기 적어도 두 종류의 염기 중 제1 염기가 합성된 제1 분자들, 및 상기 적어도 두 종류의 염기 중 제2 염기가 적어도 한 염기 합성된 제2 분자들을 포함할 수 있다.

[0014] 상기 제1 분자들과 상기 제2 분자들 간의 비율은 혼합 비율에 따를 수 있다. 상기 혼합 비율에서, 상기 적어도 두 종류의 염기들 간의 비율이 동일하거나 다를 수 있다.

[0015] 상기 적어도 두 종류의 염기들은 DNA 염기, RNA 염기, 및 핵산 유사체(Nucleic acid analogue) 중 적어도 하나

일 수 있다.

[0016] 발명의 또 다른 특징에 따른 DNA 데이터의 디코딩 방법은, 복수의 DNA 프래그멘트를 주소에 따라 분류하는 단계, 상기 분류된 복수의 DNA 프래그멘트 각각에 대해 동일한 위치에서의 염기를 분석하는 단계, 상기 분석 결과에 기초하여 염기 비율의 산포도를 산출하는 단계, 상기 염기 비율의 산포도가 적어도 두 염기들의 혼합 비율일 때, 해당 위치의 염기를 디제너레이트 염기로 결정하는 단계, 및 상기 결정된 디제너레이트 염기를 기초로 데이터를 디코딩하는 단계를 포함할 수 있다.

[0017] 상기 염기를 분석하는 단계는, 상기 복수의 DNA 프래그멘트 각각의 각 위치에 대한 염기 호출이 분석되는 단계를 포함할 수 있다.

[0018] 상기 DNA 데이터의 디코딩 방법에 있어서, 상기 염기 비율의 산포도가 적어도 두 염기의 제1 혼합 비율일 때, 해당 위치의 염기를 제1 디제너레이트 염기로 결정하고, 상기 염기 비율의 산포도가 상기 적어도 두 염기의 제2 혼합 비율일 때, 해당 위치의 염기를 제2 디제너레이트 염기로 결정하며, 상기 제1 혼합 비율과 상기 제2 혼합 비율이 서로 다르고, 상기 제1 디제너레이트 염기와 상기 제2 디제너레이트 염기가 서로 다르다.

발명의 효과

[0019] 실시 예를 통해, DNA를 이용한 디지털 데이터 저장 비용을 감소시켜 실제적으로 DNA 디지털 데이터 저장 장치 및 저장 방법, 및 디코딩 방법을 구현할 수 있다.

도면의 간단한 설명

[0020] 도 1은 실시 예에 따른 DNA 디지털 데이터 저장 장치의 일 예를 나타낸 도면이다.

도 2A-도 2C는 실시 예를 설명하기 위한 제1 실험 예로서, 도 2D의 텍스트 파일을 저장하는데 있어서, 디제너레이트 염기를 추가하여 DNA 길이를 압축하는 DNA 디지털 데이터 저장 방법을 나타낸 도면이다.

도 3A-도 3D는 제1 실험 예에 따른 DNA 디지털 데이터 저장의 구조 및 디코딩 프로세스를 나타낸 도면이다.

도 4는 훈민정음 사본의 섬네일(thumbnail) 이미지이다.

도 5는 제2 실험 예에서 데이터 조각 구조(어댑터없음) 및 오류 수정 체계를 나타낸 도면이다.

도 6A 및 도 6B는 디제너레이트 염기에서 각 염기의 호출 빈도를 나타낸 그래프이다.

도 7A-7C는 실험예 및 시뮬레이션 예들을 통해 실시 예에 따른 플랫폼의 견고성 및 확장성(robustness and scalability)을 보여주는 그래프이다.

도 8은 실시 예에 따른 비용과 종래 비용을 비교한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0021] 이하에서는 첨부한 도면을 참고로 하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명을 상세히 설명한다. 그러나, 본 발명이 본 개시에서 설명된 것에 한정되지 않고 다른 형태로 구체화될 수도 있다.

[0022] 본 개시는 해당 데이터를 저장하는 데 필요한 DNA의 길이를 압축하기 위해서, adenine(A), cytosine(C), guanine(G), 및 thymine(T) 이외의 추가 문자를 지시하는 디제너레이트 염기(degenerate bases) 또는 혼합 염기(mixed bases) 서열을 이용하여, 데이터를 DNA로 인코딩하는 장치 및 그 방법에 관한 것이다. 그러면, DNA를 이용한 데이터 저장 비용을 감소시켜, DNA를 이용한 디지털 데이터 저장 장치(이하, DNA 디지털 데이터 저장 장치)를 실제적으로 제공할 수 있다.

[0023] 디지털 데이터 저장 매체로 DNA를 사용하는 것은 두 가지 주요 이점이 있다. DNA 1 그램(gram) 당 페타바이트의 데이터를 저장할 수 있는 높은 물리적 정보 밀도와 에너지 입력 없이 수세기 동안 지속되는 내구성이다.

[0024] DNA 디지털 데이터 저장에 대한 이전 연구의 주요 목표는 데이터 오류 또는 손실을 줄이기 위한 데이터 인코딩 알고리즘을 향상시키는 것이다. 예를 들어, 오류를 발생시키는 원인으로 알려진, 높은 GC 함량과 인코딩에서의 긴 호모 폴리머(homo polymer)를 제거하는 알고리즘이 제안되어 왔다. 또한, 디코딩하는 동안 오류를 수정하거나 삭제된 데이터 조각을 복구하기 위해 DNA 디지털 데이터 저장에 대한 다양한 오류 수정 알고리즘이 개발되어 왔다. 이러한 인코딩 알고리즘에 대한 이전의 연구는 거의 모든 데이터 오류 또는 손실을 제거하여 디지털 데이

터 저장 매체로서의 DNA의 잠재력을 가속화해 왔다.

[0025] 이상적인 DNA 디지털 데이터 저장을 향한 다음 단계는 데이터 저장 비용을 줄이는 것이다. 이전 연구에 따르면 DNA 비용이 약 1/100으로 감소해야만, 백업 저장 매체로 DNA를 실용화할 수 있다. 따라서 합성된 뉴클레오티드(nucleotide) 당 저장할 수 있는 데이터의 양(정보 용량, bit/nt)을 늘리고 데이터 저장을 위한 DNA 길이를 압축함으로써 비용을 최소화해야 한다. 그러나 이전의 DNA 디지털 데이터 저장 알고리즘은 비트 인코딩에 네 가지 유형의 문자(A, C, G, T)가 사용되기 때문에, 정보 용량이 2.0 bit/nt으로 제한된다. 이전의 연구가 이론적인 최댓값에 거의 도달했기 때문에 추가 인코딩 문자를 도입하지 않으면 정보 용량이 많이 증가하지 않는다.

[0026] 실시 예는 4 개의 DNA 염기에 기초한 조합과 관련된 혼합 염기(mixed bases) 또는 디제너레이트 염기(degenerate bases)를, 정보 용량 한계를 초과하는 인코딩을 위한 추가 문자로 사용한다. 실시 예에서, 디제너레이트 염기는 추가 비용 없이 단일 염기 위치에 수많은 변형을 생성 할 수 있다.

[0027] 실시 예에서, DNA 디지털 데이터 저장 장치의 한 위치에 저장된 염기를 시퀀싱하여, 두 종류 이상의 염기 서열이 리드되는 경우, 그 위치에서의 염기 조합이 디제너레이트 염기로 검출되고, A, C, G, T 이외의 인코딩 문자에 해당되는 것으로 판단된다. 예를 들어, A, C, G, T를 포함한 총 15 개의 인코딩 문자를 사용하면, 정보 용량이 3.37 bit/nt로 향상되고, 종래 기술과 비교하여, 같은 데이터를 저장하는 데 필요한 DNA 길이를 절반으로 압축할 수 있다. 그러면, DNA 디지털 데이터 저장 비용을 절반 이상 줄일 수 있다.

[0028] 도 1은 실시 예에 따른 DNA 디지털 데이터 저장 장치의 일 예를 나타낸 도면이다.

[0029] 도 1에 도시된 바와 같이, DNA 디지털 데이터 저장 장치(1)는 복수의 메모리 셀(11)을 포함한다. 도 1에서는, 매트릭스 형태로 복수의 메모리 셀(11)이 배열되어 있는 것으로 도시되어 있으나, 발명이 이에 한정되는 것은 아니다. 메모리 셀은 튜브로 구현될 수 있고, 이 경우, DNA 디지털 데이터 저장 장치(1)는 복수의 튜브를 포함할 수 있으며, 배열 형태는 제한되지 않는다.

[0030] 복수의 메모리 셀(11) 각각에는 저장 기본 단위인 한 프래그멘트(fragment)에 대응하는 복수의 염기 서열들이 합성되어 형성된 분자들이 저장될 수 있다.

[0031] 도 2A-도 2C는 실시 예를 설명하기 위한 제1 실험 예로서, 디제너레이트 염기를 추가하여 DNA 길이를 압축하는 DNA 디지털 데이터 저장을 나타낸 도면이다.

[0032] 도 2B에 도시된 바와 같이, 디지털 데이터를 DNA로 인코딩하기 위해 A, C, G 및 T 외에 11 종류의 디제너레이트 염기를 사용할 수 있다. 그러면, 4 개에서 15 개의 문자를 기반으로한 인코딩 시스템으로의 변환이 이뤄져, 이론적으로 최대 정보 용량이 2(=$\log_2 4$) bit/nt 에서 3.90(=$\log_2 15$) bit/nt로 증가한다. 이진 데이터는 A, C, G, T 및 11 개의 추가 디제너레이트 염기로 구성된 총 15개의 인코딩 문자에 대응하는 DNA 서열(sequence)로 인코딩 될 수 있다.

[0033] 제1 실험 예에 따라 인코딩된 DNA의 길이는 추가된 인코딩 문자가 있기 때문에 A, C, G, T만 사용하는 기존의 인코딩 방법의 길이보다 줄어든다. 예를 들어, 도 2A에 도시된 바와 같이, 실시 예에서는 40bit가 11nt로 인코딩될 수 있고, 이는 "This work"로 표시되어 있다. 기존 인코딩 방법들에 따라 40bit를 인코딩한 (3)~(9) 경우보다 제1 실험 예에 따른 정보 용량(information capacity)이 대략 2배 정도 높은 것을 알 수 있다.

[0034] 디제너레이트 염기는 뉴클레오타이드의 혼합 풀(pool)로 구성된 추가 문자를 나타낸다. 인코딩된 서열 중 디제너레이트 부분의 합성 과정 동안, DNA 포스포르아미다이트(phosphoramidites)를 혼합함으로써 혼입될 수 있다. 도 2C에 도시된 바와 같이, 기판(12) 위에서 합성 순서 3'->5'에 따라 염기 'A'가 합성되고, 디제너레이트 염기 'K'를 구성하는 염기 'T'와 'G' 각각이 혼합 비율에 기초하여 합성되며, 염기 'C'가 합성되고, 디제너레이트 염기 'D'를 구성하는 'A', 'G', 'T' 각각이 혼합 비율에 기초하여 합성된다. 도 2B에서는 기판(12)에 디제너레이트 염기를 구성하는 염기들을 혼합 비율에 기초하여 혼입한다. 그러나, 발명이 이에 한정되는 것은 아니고, 다른 장치에서 디제너레이트 염기를 구성하는 염기들을 혼합 비율에 따라 혼합한 후, 혼합된 염기들을 기판(12)에 혼입할 수 있다.

[0035] 도 2C에서는 기판(12) 위의 한 컬럼(13)에서 저장 기본 단위인 한 프레그멘트에 대응하는 복수의 염기 서열들이 합성된 결과가 도시되어 있다(컬럼 방식). 그러나 복수의 염기 서열들이 합성되는 방식이 이에 한정되는 것은 아니고 잉크젯 방식에 따라 유리 기판 위에서 합성이 이뤄질 수 있다(잉크젯 방식). 컬럼 기반(column-based) 및 잉크젯 기반(inkjet-based) 올리고뉴클레오티드 합성(oligonucleotide synthesis)에 있어서, 사용된 포스포르아미다이트의 총량이 동일하기 때문에, 디제너레이트 염기는 추가 비용 없이 인코딩될 수 있다. 따라서 실시

예에서는, 동일한 양의 데이터를 저장할 때, 동일한 단위 합성 비용으로 종래의 DNA의 길이에 비해 DNA의 길이를 약 절반으로 압축할 수 있어, 데이터 저장 비용을 줄일 수 있다.

[0036] 제1 실험 예에서는, 도 2D에 도시된 854 바이트의 텍스트 파일을 DNA 서열로 인코딩한다.

[0037] 도 2D는 제1 실험 예를 위한 텍스트 파일의 일 예이다.

[0038] 데이터는 일련의 DNA 코돈(codon)으로 변형되고, 코돈은 15 개의 문자 중 3 개의 문자로 구성될 수 있다. 4 염기 이상의 호모 폴리머를 피하기 위해 코돈의 마지막 위치 서열과 코돈의 프론트 서열은 동일하지 않도록 한다.

[0039] 아래 표 1은 15 개의 문자 중 3 개의 문자로 구성된 코돈을 나타내고 있다.

[0040] [표 1]

ACA	ZTA	OYA	USA	DIA	BXA
CCA	XTA	PYA	ISA	NIA	VXA
TCA	AGA	ZYA	OSA	UIA	DXA
GCA	CGA	XYA	PSA	IIA	NXA
RCA	TGA	AKA	ZSA	OIA	UXA
YCA	GGA	CKA	XSA	PIA	IXA
MCA	RGA	TKA	ABA	ZIA	OXA
KCA	YGA	GKA	CBA	XIA	PXA
WCA	MGA	RKA	TBA	APA	ZXA
SCA	KGA	YKA	GBA	CPA	XXA
HCA	WGA	MKA	RBA	TPA	AAC
BCA	SGA	KKA	YBA	GPA	CAC
VCA	HGA	WKA	MBA	RPA	TAC
DCA	BGA	SKA	KBA	YPA	GAC
NCA	VGA	HKA	WBA	MPA	RAC
UCA	DGA	BKA	SBA	KPA	YAC
ICA	NGA	VKA	HBA	WPA	MAC
OCA	UGA	DKA	BBA	SPA	KAC
PCA	IGA	NKA	VBA	HPA	WAC
ZCA	OGA	UKA	DBA	BPA	SAC
XCA	PGA	IKA	NBA	VPA	HAC
ATA	ZGA	OKA	UBA	DPA	BAC
CTA	XGA	PKA	IBA	NPA	VAC
TTA	AYA	ZKA	OBA	UPA	DAC
GTA	CYA	XKA	PBA	IPA	NAC
RTA	TYA	ASA	ZBA	OPA	UAC
YTA	GYA	CSA	XBA	PPA	IAC
MTA	RYA	TSA	AIA	ZPA	OAC
KTA	YYA	GSA	CIA	XPA	PAC
WTA	MYA	RSA	TIA	AXA	ZAC
STA	KYA	YSA	GIA	CXA	XAC
HTA	WYA	MSA	RIA	TXA	ATC
BTA	SYA	KSA	YIA	GXA	CTC
VTA	HYA	WSA	MIA	RXA	TTC
DTA	BYA	SSA	KIA	YXA	GTC
NTA	VYA	HSA	WIA	MXA	RTC
UTA	DYA	BSA	SIA	KXA	YTC
ITA	NYA	VSA	HIA	WXA	MTC
OTA	UYA	DSA	BIA	SXA	KTC
PTA	IYA	NSA	VIA	HXA	WTC
STC	KRC	YWC	GUC	CZC	XAT
HTC	WRC	MWC	RUC	TZC	ACT
BTC	SRC	KWC	YUC	GZC	CCT
VTC	HRC	WWC	MUC	RZC	TCT

[0041]

[0042]

DTC	BRC	SWC	KUC	YZC	GCT
NTC	VRC	HWC	WUC	MZC	RCT
UTC	DRC	BWC	SUC	KZC	YCT
ITC	NRC	VWC	HUC	WZC	MCT
OTC	URC	DWC	BUC	SZC	KCT
PTC	IRC	NWC	VUC	HZC	WCT
ZTC	ORC	UWC	DUC	BZC	SCT
XTC	PRC	IWC	NUC	VZC	HCT
AGC	ZRC	OWC	UUC	DZC	BCT
CGC	XRC	PWC	IUC	NZC	VCT
TGC	AKC	ZWC	OUC	UZC	DCT
GGC	CKC	XWC	PUC	IZC	NCT
RGC	TKC	ADC	ZUC	OZC	UCT
YGC	GKC	CDC	XUC	PZC	ICT
MGC	RKC	TDC	APC	ZZC	OCT
KGC	YKC	GDC	CPC	XZC	PCT
WGC	MKC	RDC	TPC	AAT	ZCT
SGC	KKC	YDC	GPC	CAT	XCT
HGC	WKC	MDC	RPC	TAT	AGT
BGC	SKC	KDC	YPC	GAT	CGT
VGC	HKC	WDC	MPC	RAT	TGT
DGC	BKC	SDC	KPC	YAT	GGT
NGC	VKC	HDC	WPC	MAT	RGT
UGC	DKC	BDC	SPC	KAT	YGT
IGC	NKC	VDC	HPC	WAT	MGT
OGC	UKC	DDC	BPC	SAT	KGT
PGC	IKC	NDC	VPC	HAT	WGT
ZGC	OKC	UDC	DPC	BAT	SGT
XGC	PKC	IDC	NPC	VAT	HGT
ARC	ZKC	ODC	UPC	DAT	BGT
CRC	XKC	PDC	IPC	NAT	VGT
TRC	AWC	ZDC	OPC	UAT	DGT
GRC	CWC	XDC	PPC	IAT	NGT
RRC	TWC	AUC	ZPC	OAT	UGT
YRC	GWC	CUC	XPC	PAT	IGT
MRC	RWC	TUC	AZC	ZAT	OGT
PGT	IMT	NVT	VOT	HAG	WTG
ZGT	OMT	UVT	DOT	BAG	STG
XGT	PMT	IVT	NOT	VAG	HTG
ART	ZMT	OVT	UOT	DAG	BTG
CRT	XMT	PVT	IOT	NAG	VTG
TRT	AST	ZVT	OOT	UAG	DTG
GRT	CST	XVT	POT	IAG	NTG
RRT	TST	AUT	ZOT	OAG	UTG
YRT	GST	CUT	XOT	PAG	ITG
MRT	RST	TUT	AXT	ZAG	OTG
KRT	YST	GUT	CXT	XAG	PTG

[0043]

WRT	MST	RUT	TXT	ACG	ZTG
SRT	KST	YUT	GXT	CCG	XTG
HRT	WST	MUT	RXT	TCG	AYG
BRT	SST	KUT	YXT	GCG	CYG
VRT	HST	WUT	MXT	RCG	TYG
DRT	BST	SUT	KXT	YCG	GYG
NRT	VST	HUT	WXT	MCG	RYG
URT	DST	BUT	SXT	KCG	YYG
IRT	NST	VUT	HXT	WCG	MYG
ORT	UST	DUT	BXT	SCG	KYG
PRT	IST	NUT	VXT	HCG	WYG
ZRT	OST	UUT	DXT	BCG	SYG
XRT	PST	IUT	NXT	VCG	HYG
AMT	ZST	OUT	UXT	DCG	BYG
CMT	XST	PUT	IXT	NCG	VYG
TMT	AVT	ZUT	OXT	UCG	DYG
GMT	CVT	XUT	PXT	ICG	NYG
RMT	TVT	AOT	ZXT	OCG	UYG
YMT	GVT	COT	XXT	PCG	IYG
MMT	RVT	TOT	AAG	ZCG	OYG
KMT	YVT	GOT	CAG	XCG	PYG
WMT	MVT	ROT	TAG	ATG	ZYG
SMT	KVT	YOT	GAG	CTG	XYG
HMT	WVT	MOT	RAG	TTG	AMG
BMT	SVT	KOT	YAG	GTG	CMG
VMT	HVT	WOT	MAG	RTG	TMG
DMT	BVT	SOT	KAG	YTG	GMG
NMT	VVT	HOT	WAG	MTG	RMG
UMT	DVT	BOT	SAG	KTG	YMG
MMG	RHG	TOG	ACR	ZTR	OIR
KMG	YHG	GOG	CCR	XTR	PIR
WMG	MHG	ROG	TCR	AYR	ZIR
SMG	KHG	YOG	GCR	CYR	XIR
HMG	WHG	MOG	RCR	TYR	AAY
BMG	SHG	KOG	YCR	GYR	CAY
VMG	HHG	WOG	MCR	RYR	TAY
DMG	BHG	SOG	KCR	YYR	GAY
NMG	VHG	HOG	WCR	MYR	RAY
UMG	DHG	BOG	SCR	KYR	YAY
IMG	NHG	VOG	HCR	WYR	MAY
OMG	UHG	DOG	BCR	SYR	KAY
PMG	IHG	NOG	VCR	HYR	WAY
ZMG	OHG	UOG	DCR	BYR	SAY
XMG	PHG	IOG	NCR	VYR	HAY
AWG	ZHG	OOG	UCR	DYR	BAY
CWG	XHG	POG	ICR	NYR	VAY
TWG	AIG	ZOG	OCR	UYR	DAY

GWG	CIG	XOG	PCR	IYR	NAY
RWG	TIG	AZG	ZCR	OYR	UAY
YWG	GIG	CZG	XCR	PYR	IAY
MWG	RIG	TZG	ATR	ZYR	OAY
KWG	YIG	GZG	CTR	XYR	PAY
WWG	MIG	RZG	TTR	AIR	ZAY
SWG	KIG	YZG	GTR	CIR	XAY
HWG	WIG	MZG	RTR	TIR	AGY
BWG	SIG	KZG	YTR	GIR	CGY
VWG	HIG	WZG	MTR	RIR	TGY
DWG	BIG	SZG	KTR	YIR	GGY
NWG	VIG	HZG	WTR	MIR	RGY
UWG	DIG	BZG	STR	KIR	YGY
IWG	NIG	VZG	HTR	WIR	MGY
OWG	UIG	DZG	BTR	SIR	KGY
PWG	IIG	NZG	VTR	HIR	WGY
ZWG	OIG	UZG	DTR	BIR	SGY
XWG	PIG	IZG	NTR	VIR	HGY
AHG	ZIG	OZG	UTR	DIR	BGY
CHG	XIG	PZG	ITR	NIR	VGY
THG	AOG	ZZG	OTR	UIR	DGY
GHG	COG	XZG	PTR	IIR	NGY
UGY	DUY	BGM	SPM	KCK	YOK
IGY	NUY	VGM	HPM	WCK	MOK
OGY	UUY	DGM	BPM	SCK	KOK
PGY	IUY	NGM	VPM	HCK	WOK
ZGY	OUY	UGM	DPM	BCK	SOK
XGY	PUY	IGM	NPM	VCK	HOK
ARY	ZUY	OGM	UPM	DCK	BOK
CRY	XUY	PGM	IPM	NCK	VOK
TRY	ATM	ZGM	OPM	UCK	DOK
GRY	CTM	XGM	PPM	ICK	NOK
RRY	TTM	AKM	ZPM	OCK	UOK
YRY	GTM	CKM	XPM	PCK	IOK
MRY	RTM	TKM	AAK	ZCK	OOK
KRY	YTM	GKM	CAK	XCK	POK
WRY	MTM	RKM	TAK	AMK	ZOK
SRY	KTM	YKM	GAK	CMK	XOK
HRY	WTM	MKM	RAK	TMK	ACW
BRY	STM	KKM	YAK	GMK	CCW
VRY	HTM	WKM	MAK	RMK	TCW
DRY	BTM	SKM	KAK	YMK	GCW
NRY	VTM	HKM	WAK	MMK	RCW
URY	DTM	BKM	SAK	KMK	YCW
IRY	NTM	VKM	HAK	WMK	MCW
ORY	UTM	DKM	BAK	SMK	KCW
PRY	ITM	NKM	VAK	HMK	WCW

[0044]

[0045]

ZRY	OTM	UKM	DAK	BMK	SCW
XRY	PTM	IKM	NAK	VMK	HCW
AUY	ZTM	OKM	UAK	DMK	BCW
CUY	XTM	PKM	IAK	NMK	VCW
TUY	AGM	ZKM	OAK	UMK	DCW
GUY	CGM	XKM	PAK	IMK	NCW
RUY	TGM	APM	ZAK	OMK	UCW
YUY	GGM	CPM	XAK	PMK	ICW
MUY	RGM	TPM	ACK	ZMK	OCW
KUY	YGM	GPM	CCK	XMK	PCW
WUY	MGM	RPM	TCK	AOK	ZCW
SUY	KGM	YPM	GCK	COK	XCW
HUY	WGM	MPM	RCK	TOK	AGW
BUY	SGM	KPM	YCK	GOK	CGW
VUY	HGM	WPM	MCK	ROK	TGW
GGW	CXW	XAS	PWS	IGH	NTV
RGW	TXW	ATS	ZWS	OGH	UTV
YGW	GXW	CTS	XWS	PGH	ITV
MGW	RXW	TTS	AZS	ZGH	OTV
KGW	YXW	GTS	CZS	XGH	PTV
WGW	MXW	RTS	TZS	AAB	ZTV
SGW	KXW	YTS	GZS	CAB	XTV
HGW	WXW	MTS	RZS	TAB	ACD
BGW	SXW	KTS	YZS	GAB	CCD
VGW	HXW	WTS	MZS	RAB	TCD
DGW	BXW	STS	KZS	YAB	GCD
NGW	VXW	HTS	WZS	MAB	RCD
UGW	DXW	BTS	SZS	KAB	YCD
IGW	NXW	VTS	HZS	WAB	MCD
OGW	UXW	DTS	BZS	SAB	KCD
PGW	IXW	NTS	VZS	HAB	WCD
ZGW	OXW	UTS	DZS	BAB	SCD
XGW	PXW	ITS	NZS	VAB	HCD
ASW	ZXW	OTS	UZS	DAB	BCD
CSW	XXW	PTS	IZS	NAB	VCD
TSW	AAS	ZTS	OZS	UAB	DCD
GSW	CAS	XTS	PZS	IAB	NCD
RSW	TAS	AWS	ZZS	OAB	UCD
YSW	GAS	CWS	XZS	PAB	ICD
MSW	RAS	TWS	AGH	ZAB	OCD
KSW	YAS	GWS	CGH	XAB	PCD
WSW	MAS	RWS	TGH	ATV	ZCD
SSW	KAS	YWS	GGH	CTV	XCD
HSW	WAS	MWS	RGH	TTV	ACU
BSW	SAS	KWS	YGH	GTV	CCU
VSW	HAS	WWS	MGH	RTV	TCU
DSW	BAS	SWS	KGH	YTV	GCU

[0046]

NSW	VAS	HWS	WGH	MTV	RCU
USW	DAS	BWS	SGH	KTV	YCU
ISW	NAS	VWS	HGH	WTV	MCU
OSW	UAS	DWS	BGH	STV	KCU
PSW	IAS	NWS	VGH	HTV	WCU
ZSW	OAS	UWS	DGH	BTV	SCU
XSW	PAS	IWS	NGH	VTV	HCU
AXW	ZAS	OWS	UGH	DTV	BCU
VCU	HYU	WAI	MRI	RTO	TKO
DCU	BYU	SAI	KRI	YTO	GKO
NCU	VYU	HAI	WRI	MTO	RKO
UCU	DYU	BAI	SRI	KTO	YKO
ICU	NYU	VAI	HRI	WTO	MKO
OCU	UYU	DAI	BRI	STO	KKO
PCU	IYU	NAI	VRI	HTO	WKO
ZCU	OYU	UAI	DRI	BTO	SKO
XCU	PYU	IAI	NRI	VTO	HKO
ATU	ZYU	OAI	URI	DTO	BKO
CTU	XYU	PAI	IRI	NTO	VKO
TTU	AIU	ZAI	ORI	UTO	DKO
GTU	CIU	XAI	PRI	ITO	NKO
RTU	TIU	AGI	ZRI	OTO	UKO
YTU	GIU	CGI	XRI	PTO	IKO
MTU	RIU	TGI	AUI	ZTO	OKO
KTU	YIU	GGI	CUI	XTO	PKO
WTU	MIU	RGI	TUI	AGO	ZKO
STU	KIU	YGI	GUI	CGO	XKO
HTU	WIU	MGI	RUI	TGO	APO
BTU	SIU	KGI	YUI	GGO	CPO
VTU	HIU	WGI	MUI	RGO	TPO
DTU	BIU	SGI	KUI	YGO	GPO
NTU	VIU	HGI	WUI	MGO	RPO
UTU	DIU	BGI	SUI	KGO	YPO
ITU	NIU	VGI	HUI	WGO	MPO
OTU	UIU	DGI	BUI	SGO	KPO
PTU	IIU	NGI	VUI	HGO	WPO
ZTU	OIU	UGI	DUI	BGO	SPO
XTU	PIU	IGI	NUI	VGO	HPO
AYU	ZIU	OGI	UUI	DGO	BPO
CYU	XIU	PGI	IUI	NGO	VPO
TYU	AAI	ZGI	OUI	UGO	DPO
GYU	CAI	XGI	PUI	IGO	NPO
RYU	TAI	ARI	ZUI	OGO	UPO
YYU	GAI	CRI	XUI	PGO	IPO
MYU	RAI	TRI	ATO	ZGO	OPO
KYU	YAI	GRI	CTO	XGO	PPO
WYU	MAI	RRI	TTO	AKO	ZPO

SYU	KAI	YRI	GTO	CKO	XPO
AAP	ZCP	OOP	UGZ	DXZ	BTX
CAP	XCP	POP	IGZ	NXZ	VTX
TAP	AMP	ZOP	OGZ	UXZ	DTX
GAP	CMP	XOP	PGZ	IXZ	NTX
RAP	TMP	ACZ	ZGZ	OXZ	UTX
YAP	GMP	CCZ	XGZ	PXZ	ITX
MAP	RMP	TCZ	ASZ	ZXZ	OTX
KAP	YMP	GCZ	CSZ	XXZ	PTX
WAP	MMP	RCZ	TSZ	AAX	ZTX
SAP	KMP	YCZ	GSZ	CAX	XTX
HAP	WMP	MCZ	RSZ	TAX	AWX
BAP	SMP	KCZ	YSZ	GAX	CWX
VAP	HMP	WCZ	MSZ	RAX	TWX
DAP	BMP	SCZ	KSZ	YAX	GWX
NAP	VMP	HCZ	WSZ	MAX	RWX
UAP	DMP	BCZ	SSZ	KAX	YWX
IAP	NMP	VCZ	HSZ	WAX	MWX
OAP	UMP	DCZ	BSZ	SAX	KWX
PAP	IMP	NCZ	VSZ	HAX	WWX
ZAP	OMP	UCZ	DSZ	BAX	SWX
XAP	PMP	ICZ	NSZ	VAX	HWX
ACP	ZMP	OCZ	USZ	DAX	BWX
CCP	XMP	PCZ	ISZ	NAX	VWX
TCP	AOP	ZCZ	OSZ	UAX	DWX
GCP	COP	XCZ	PSZ	IAX	NWX
RCP	TOP	AGZ	ZSZ	OAX	UWX
YCP	GOP	CGZ	XSZ	PAX	IWX
MCP	ROP	TGZ	AXZ	ZAX	OWX
KCP	YOP	GGZ	CXZ	XAX	PWX
WCP	MOP	RGZ	TXZ	ATX	ZWX
SCP	KOP	YGZ	GXZ	CTX	XWX
HCP	WOP	MGZ	RXZ	TTX	AZX
BCP	SOP	KGZ	YXZ	GTX	CZX
VCP	HOP	WGZ	MXZ	RTX	TZX
DCP	BOP	SGZ	KXZ	YTX	GZX
NCP	VOP	HGZ	WXZ	MTX	RZX
UCP	DOP	BGZ	SXZ	KTX	YZX
ICP	NOP	VGZ	HXZ	WTX	MZX
OCP	UOP	DGZ	BXZ	STX	KZX
PCP	IOP	NGZ	VXZ	HTX	WZX
SZX	BZX	DZX	UZX	OZX	ZZX
HZX	VZX	NZX	IZX	PZX	XZX

[0047]

[0048] 도 3A-도 3D는 제1 실험 예에 따른 DNA 디지털 데이터 저장의 구조 및 디코딩 프로세스를 나타낸 도면이다.

[0049] 도 3A는 DNA 프래그멘트(fragment)의 디자인 구조이고, 도 3B는 DNA 프래그멘트가 NGS(Next Generation Sequencing)에 의해 분석되는 것을 나타내고 있다. 주소에 따라 분류를 한 후에, DNA 디지털 데이터 장치의 각 컬럼에서 동일한 위치에서의 DNA 염기에 기초한 문자 분배가 분석되고, 그 결정에 따라 디제너레이트 염기가 디코딩될 수 있다. 예를 들어, 도 3C에 도시된 바와 같이, 동일한 위치에서의 염기를 분석하고, 분석한 결과에 기초하여 염기 비율의 산포도(scatter plot)를 산출하며, 염기 비율의 산포도에 따라 디제너레이트 염기가 결정될 수 있다. 도 3D에서, 전체 프레그멘트에 대한 특정 평균 적용 범위에서 결정된 DNA 염기의 에러율이 도시되어 있다. 도 3D의 그래프에서, 표준 편차(s.d.)는 무작위 샘플링을 5 번 반복함으로써 얻어졌고, 오차 막대는 s.d.를 나타낸다.

[0050] 도 3A에 도시된 바와 같이, 부호화된 정보는 42nt의 단편으로 분할되고, 길이 3nt의 염기로 구성된 주소가 할당된다. 주소를 나타내는 3nt의 염기는 디제너레이트 염기를 포함하지 않고, 그 예는 아래 표 2과 같다.

[0051] [표 2]

1	ACA	13	AAC	25	AAT	37	AAG
2	CCA	14	CAC	26	CAT	38	CAG
3	TCA	15	TAC	27	TAT	39	TAG
4	GCA	16	GAC	28	GAT	40	GAG
5	ATA	17	ATC	29	ACT	41	ACG
6	CTA	18	CTC	30	CCT	42	CCG
7	TTA	19	TTC	31	TCT	43	TCG
8	GTA	20	GTC	32	GCT	44	GCG
9	AGA	21	AGC	33	AGT	45	ATG
10	CGA	22	CGC	34	CGT	46	CTG
11	TGA	23	TGC	35	TGT	47	TTG
12	GGA	24	GGC	36	GGT	48	GTG

[0052]

[0053] 각 프래그멘트는 증폭 및 시퀀싱을 위해 2 개의 어댑터 (5 '및 3'말단 각각 에 20nt)로 보충되며, 전체 단편의 길이는 85nt이다. 제1 실험 예에서, 45 개의 DNA 프래그멘트가 컬럼-기초 올리고뉴클레오티드 합성기(column-based oligonucleotide synthesizer)에 의해 합성되었다. 어댑터를 제외한 전체 뉴클레오티드 합성에서 인코딩된 비트의 수를 고려하면, 3.37bit/nt의 정보 용량이 제1 실험 예에서 제공될 수 있다. 평균 800 분자로 구성된 합성 DNA 라이브러리는 설계된 어댑터에 의해 증폭되고 "Illumina MiniSeq"에 의해 시퀀싱될 수 있다.

[0054] 원 NGS 데이터는 설계된 길이로 필터링되고 주소별로 분류된다. 도 3B에 도시된 바와 같이, 분류된 다음 복제된 리드(read)가 제거되고 프레그멘트의 각 위치에 대한 염기 호출(call)이 분석된다. 분석되는 뉴클레오타이드의 중간 비율은 합성 동안의 커플링 효율이 성장하는 올리고뉴클레오타이드의 유형 및 위치에 따라 각각의 염기마다 변하기 때문에 일관되게 동일하지는 않다.

[0055] 그러나, 동일한 위치에서 분석한 서열에서 A:C:G:T의 비율을 산포도로 관찰 할 때, 도 3C에 도시된 바와 같이, 전체 분포는 15 개의 클러스터로 나누어지고, 그 중 2 개 이상의 염기들이 혼합 비율에 따라 구성된 11 개의 클러스터가 디제너레이트 염기로 간주된다. 특정 뉴클레오타이드의 지배적인 비율을 갖는 나머지 4 클러스터는 순수한 염기 서열로 간주된다. 이와 같이, 제1 실험 예에서, 원 NGS 데이터에서 원 데이터를 성공적으로 복구할 수 있다.

[0056] 도 3D에서는, 평균 10 회 무작위로 평균 250x 커버리지(coverage)로 다운 샘플링 한 10 건에서, 데이터를 복구할 수 있는 점을 볼 수 있다. 1x 커버리지는 설계된 모든 DNA에 대해서 NGS로 DNA의 종류를 검출하여 데이터를 읽는 것을 의미한다. 원 NGS 데이터(예를 들어, 3600x 커버리지)에서 250x 커버리지를 랜덤하게 다운 샘플링하여 인코딩된 원 데이터와 비교한 결과 오차가 없었고, 이 다운 샘플링을 10번 반복한 다운 샘플링 한 10건 모두에서 오차가 발생하지 않았다. 정보평균 NGS 커버리지가 250x보다 낮으면 인코딩 문자들의 클러스터들 간의 교차점이 증가하기 때문에 에러율이 증가할 수 있다.

[0057] 도 4는 훈민정음 사본의 섬네일(thumbnail) 이미지이다.

[0058] 실시 예에 따른 DNA 디지털 데이터 저장 플랫폼의 확장성을 보여주기 위해 제2 실험 예에서는, 풀링된 올리고뉴클레오타이드 합성(pooled oligonucleotide synthesis) 방법을 사용하여 4503 개의 DNA 프래그멘트에 훈민정음 사본의 섬네일 이미지 데이터인 135.4Kbytes를 저장하였다.

[0059] 도 5는 제2 실험 예에서 데이터 조각 구조(어댑터없음) 및 오류 수정 체계를 나타낸 도면이다.

[0060] 도 5에 도시된 바와 같이, 높은 복잡성으로 올리고뉴클레오타이드 풀을 합성 및 증폭할 때 발생할 수 있는 오차와 증폭 바이어스에 대처하기 위해 Reed-Solomon 기반 중복(Reed-Solomon based redundancy)을 추가했다.

[0061] 도 5에서 도시된 실시 예에서는 디제너레이트 염기 'W'와 'S'를 추가하여, 데이터를 인코딩하였다. 표 3은 디제

너레이트 염기 W 및 S를 포함한 코돈(codon) 테이블이다.

[0062] [표 3]

Data	Codon	Data	Codon	Data	Codon	Data	Codon
1	ACA	33	TTC	65	WGT	97	ACW
2	CCA	34	GTC	66	SGT	98	CCW
3	TCA	35	WTC	67	AST	99	TCW
4	GCA	36	STC	68	CST	100	GCW
5	WCA	37	AGC	69	TST	101	WCW
6	SCA	38	CGC	70	GST	102	SCW
7	ATA	39	TGC	71	WST	103	AGW
8	CTA	40	GGC	72	SST	104	CGW
9	TTA	41	WGC	73	AAG	105	TGW
10	GTA	42	SGC	74	CAG	106	GGW
11	WTA	43	AWC	75	TAG	107	WGW
12	STA	44	CWC	76	GAG	108	SGW
13	AGA	45	TWC	77	WAG	109	ASW
14	CGA	46	GWC	78	SAG	110	CSW
15	TGA	47	WWC	79	ACG	111	TSW
16	GGA	48	SWC	80	CCG	112	GSW
17	WGA	49	AAT	81	TCG	113	WSW
18	SGA	50	CAT	82	GCG	114	SSW
19	ASA	51	TAT	83	WCG	115	AAS
20	CSA	52	GAT	84	SCG	116	CAS
21	TSA	53	WAT	85	ATG	117	TAS
22	GSA	54	SAT	86	CTG	118	GAS
23	WSA	55	ACT	87	TTG	119	WAS
24	SSA	56	CCT	88	GTG	120	SAS
25	AAC	57	TCT	89	WTG	121	ATS
26	CAC	58	GCT	90	STG	122	CTS
27	TAC	59	WCT	91	AWG	123	TTS
28	GAC	60	SCT	92	CWG	124	GTS
29	WAC	61	AGT	93	TWG	125	WTS
30	SAC	62	CGT	94	GWG	126	STS
31	ATC	63	TGT	95	WWG	127	AWS
32	CTC	64	GGT	96	SWG	128	CWS

[0063]

[0064] 도 5에 도시된 바와 같아, 인코딩된 데이터 111nt는 표 3의 코돈에 기초하여 37bit의 디지털 데이터로 디코딩된다. 예를 들어, 디코딩된 원정보(original information)가 에러를 가지고 있을 때, 중복(RS)에 기초하여 에러가 수정되어, 원정보가 에러 없이 복구될 수 있다.

[0065] 제2 실험 예에서도, 제1 실험 예와 동일하게, 원 NGS 데이터에서 250x 커버리지를 랜덤하게 다운 샘플링한 결과를 10번 반복하였다. 그 결과 평균 250x 커버리지(average coverage to 250x)에서 원 데이터를 오차 없이 복구하여, 정보 용량 2.0 bit/nt이 달성되었다.

[0066] 제1 및 제2 실험 예에서의 플랫폼은 순 정보 용량, 입력 데이터, 올리고 (oligo) 수, 최소 커버리지 및 물리적 밀도 측면에서 Erlich and Zielinski에 의한 플랫폼과 더 자세히 분석되고 비교될 수 있다.

[0067] 표 4는 제1 및 제2 실험 예와 Erlich and Zielinski간의 순 정보 용량(Net Information capacity (bi/nt), 입력 데이터의 사이즈(Input data), 완전 복구(Full recovery) 여부, 올리고 수(Number of oligos), 최소 NGS 커버리지(평균)(Minimum NGS coverage (average)) 및 물리적 밀도(Physical density (Pbytes/g)) 측면을 비교한

표이다.

[0068] [표 4]

	Erilch and Zielinski	This work	
Net Information capacity (bit/nt)	1.57	2	3.37
Input data	2.15 Mbyte	135.4Kbyte	854byte
Full recovery	Yes	Yes	Yes
Number of oligos	72,000	4503	45
Minimum NGS coverage (average)	10x	250x	200x
Physical density (Pbytes/g)	214	772	485

[0069]

[0070] 단일 디자인 프래그멘트에서 다수의 올리고뉴클레오타이드 변이체를 합성하였지만, 제1 및 제2 실험 예 각각에서의 데이터 복구에 필요한 디자인 당 올리고뉴클레오타이드 분자 수는 438개 및 800개이다. 이는 종래 Erlich and Zielinski에서 데이터 해독에 올리고뉴클레오타이드 분자 1300개가 사용된 것과 비교해 개선된 점 중 하나이다.

[0071] 또한, 제1 및 제2 실험에서 입증된 순 정보 용량과 물리적 밀도 역시 종래 Erlich and Zielinski에 비해서 개선된 것을 알 수 있다. 순정보 용량은 뉴클레오타이드(nt) 당 저장 가능한 비트(bit) 수로 정의되고, 물리적 밀도는 데이터를 복구하기 위해서 필요한 동일 분자수를 실험적으로 산출하고, 산출된 분자수에 기초해서 단위 무게(g) 당 저장할 수 있는 데이터의 크기(Pbyte)를 나타낸다.

[0072] 이와 같이, 제1 및 제2 실험에 기초한 데이터에 근거하여 다양한 유형의 디제너레이트 염기가 대규모로 사용되었을 때 데이터 복구의 가능성을 시뮬레이션했다.

[0073] 도 6A 및 도 6B는 디제너레이트 염기에서 각 염기의 호출 빈도를 나타낸 그래프이다.

[0074] 디제너레이트 염기 W를 인코딩하기 위해서 염기 A 및 T를 동일한 비율로 혼합하여도, 디코딩에서 호출되는 비율은 동일하지 않다. 도 6A에 도시된 바와 같이, 50개의 디제너레이트 염기 W에 대해서 염기 A의 호출 비율은 0.34145이고, 염기 T의 호출 비율은 0.64461이며, 염기 A 및 T 각각의 호출 빈도가 이항 분포를 따른다.

[0075] 마찬가지로, 디네너레이트 염기 S를 인코딩하기 위해서 염기 C 및 G를 동일한 비율로 혼합하여도, 디코딩에서 호출되는 비율은 동일하지 않다. 도 6B에 도시된 바와 같이, 50개의 디제너레이트 염기 S에 대해서 염기 C의 호출 비율은 0.46355이고, 염기 G의 호출 비율은 0.51541이며, 염기 C 및 G 각각의 호출 빈도가 이항 분포를 따른다.

[0076] 각 디제너레이트 염기가 포함하는 염기 각각의 호출 빈도도 이항 분포를 따르기 때문에, 실시 예에 따른 플랫폼은 몬테카를로(Monte-Carlo) 시뮬레이션을 통해 모델링될 수 있다. 몬테카를로 시뮬레이션은, 확률적 분포가 산출되었을 때, 전산적으로 산출된 확률적 분포를 변수 생성하여 확인하는 일반적인 모델링 기법이다.

[0077] 몬테카를로 시뮬레이션을 통해 모델링되는 과정은, 특정 확률분포(이항분포)에 빈도가 따르는 것에 기반하여, 1) 디제너레이트 염기를 포함하여 인코딩된 랜덤 데이터를 생성, 2) 랜덤 데이터에 대한 염기 호출분포를 이항분포에 기반하여 생성, 3) 생성된 염기 호출분포와 실제 염기 호출분포 간의 에러율 생성(또는, 데이터 복구가 가능한지 확인)의 동작 1-3을 반복하는 것을 포함할 수 있다.

[0078] 도 7A-7C는 실험예 및 시뮬레이션 예들을 통해 실시 예에 따른 플랫폼의 견고성 및 확장성(robustness and scalability)을 보여주는 그래프이다.

[0079] 도 7A는 프래그멘트의 리드(read) 커버리지에 따라 염기쌍 당 에러율을 나타낸 그래프이다. 도 7A의 그래프는 무작위로 그리고 일률적으로 리드들이 생성된 결과에 기초하거나, 실험 데이터에 대해서 샘플링 된 것을 기초로 할 수 있다.

[0080] 도 7A에서, 프레그멘트 당 커버리지 수에 따라 염기 당 에러율이 도시되어 있다. 이 때, 에러율은 염기들의 에러율들을 평균하여 산출될 수 있다. 도 7A에 도시된 바와 같이, 프레그멘트 당 커버리지의 수가 증가할수록 염기 당 에러율이 감소하는 것을 실험 예와 시뮬레이션 예들에서 알 수 있다.

[0081] 2개의 디제너리에트 염기 W 및 S를 포함하는 실험 예의 결과와 2개의 디제너리에트 염기 W 및 S를 포함하는 시뮬레이션의 결과가 유사한 패턴을 보이고 있다. 이에, 디제너레이트 염기들(R, Y, M, K, S, W, H, B, V, D, N)을 포함하는 15개의 문자로 인코딩한 데이터에 대한 시뮬레이션 결과에 기초하면, 그 실험 예도 유사한 결과가 도출될 수 있음을 예측할 수 있다.

[0082] 또한, 12개의 디제너레이트 염기들과 4 개의 디제너레이트 염기들(H, B, V, D, N)을 포함하는 21개의 문자로 인코딩한 데이터에 대한 시뮬레이션 결과에 기초하면, 그 실험 예도 유사한 결과가 도출될 수 있음을 예측할 수 있다. 12개의 디제너레이트 염기들 각각은 혼합 염기들의 혼합비를 다르게 하여 생성될 수 있다. 예를 들어, 디제너레이트 염기들(R, Y, M, K, S, W) 각각은 염기들의 혼합비가 3:7인 디제너레이트 염기와 혼합비가 7:3인 디제너레이트 염기를 포함할 수 있다. 구체적으로, A:T = 3:7의 경우 W1을 지정하고, A:T = 7:3의 경우 W2를 지정할 수 있다.

[0083] 도 7B는 실험을 통해 PCR 바이어스에 의해 프레그멘트가 호출되는 빈도를 나타낸 그래프이다. 도 7B에 도시된 바와 같이, 프레그멘트의 호출 빈도는 고르지 않은 프로파일(profile)을 나타낸다. 도 7B에서, NGS 커버리지 횟수에 따라 리드되는 프레그멘트의 호출 빈도가 나타나 있다. 예를 들어, NGS 커버리지 횟수가 100x, 500x, 및 1000x일 때의 프래그멘트의 호출 빈도 분포가 도시되어 있다. 즉, NGS 커버리지가 시행될 때마다, 모든 프레그멘트가 리드되지 않고, 리드되지 않는 프래그멘트가 발생한다. 또한, NGS 커버리지 횟수가 증가하여도 프래그멘트의 호출 빈도 분포가 개선되는 것은 아니다. 도 7B에서 적색 라인은 음 이항 적합(negative binomial fit)을 나타내고 있다.

[0084] 도 7C는 프레그멘트가 호출되는 빈도가 고르지 않은 것을 적용할 때, 전체 프래그멘트에 대한 NGS 커버리지 횟수에 따른 염기쌍 당 에러율을 나타내고 있다. 도 7C의 그래프에서, 실험 결과의 표준 편차(s.d.)는 무작위 샘플링을 5 번 반복함으로써 얻어졌고, 오차 막대는 s.d. 를 나타낸다.

[0085] 도 7A에 도시된 프래그멘트의 커버리지 당 염기 에러율에 도 7B에 도시된 프래크멘트의 커버리지에 따른 호출 빈도를 적용하여 도 7C에 도시된 전체 프래그멘트에 대한 NGS 커버리지 횟수에 따른 염기당 에러율을 시뮬레이션 할 수 있다.

[0086] 다양한 유형의 디제너레이트 염기를 사용하면 에러율이 증가하지만 에러율은 NGS 커버리지를 증가함에 따라 감소한다. 시뮬레이션에서 1300x 이상의 NGS 커버리지가 주어지면, 10% Reed-Solomon redundancy를 가지는 100 메가 바이트의 정보를 완벽하게 디코딩 할 수 있다.

[0087] 본 시뮬레이션에서, 디제너레이트 염기를 이용하여 이전에 보고된 DNA 디지털 데이터 저장 시스템의 순 정보 용량과 물리적 밀도를 실험적으로 두 배로 늘릴 수 있는 점을 알 수 있다. 또한, 시뮬레이션을 통해 플랫폼의 확장성을 보여주었다. 도입된 플랫폼이 많은 양의 NGS를 필요로 하지만, 시퀀싱기술은 진화 속도가 더 빠르며 현재의 DNA 시퀀싱 가격은 DNA 디지털데이터 스토리지에 사용되는 염기 당 합성 가격보다 약 50,000 배 더 낮다. 이로부터, 실시 예에 따른 플랫폼이 2000x의 NGS 커버리지를 사용한다고 해도, 시퀀싱 비용은 합성 비용의 4%에 불과하다. 실시 예는 이전 보고서(Erlich and Zielinski)의 정보 용량보다 두 배 이상 높은 DNA 디지털 데이터 저장 장치의 정보 용량을 제공할 수 있다.

[0088] 도 8은 실시 예에 따른 비용과 종래 비용을 비교한 그래프이다.

[0089] 도 8에서, 인코딩 문자가 4개인 경우의 비용은 Erlich and Zielinski 의 셋업을 기준으로 산출하였고, 실시 예에 따른 인코딩 문자가 15개 및 20개인 경우의 비용은 Erlich and Zielinski의 셋업 길이에 맞추어, 프래그멘트의 길이가 200nt이고, 어드레스의 길이는 12nt이며, 아답터(adapter)가 양단에 20nt으로 붙어있으며, 10%의 reed-solomon error correction 을 넣어주는 것으로 디자인할 때의 비용이다.

[0090] 도 8에 도시된 바와 같이, DNA 시퀀싱가격이 종래에 비해 증가하더라도, 풀-베이스(pool-based) 올리고뉴클레오타이드 합성기가 디제너레이트 염기 합성할 때, DNA 디지털 데이터 저장 비용을 절반 이상 줄일 수 있는 효과를 제공할 수 있다.

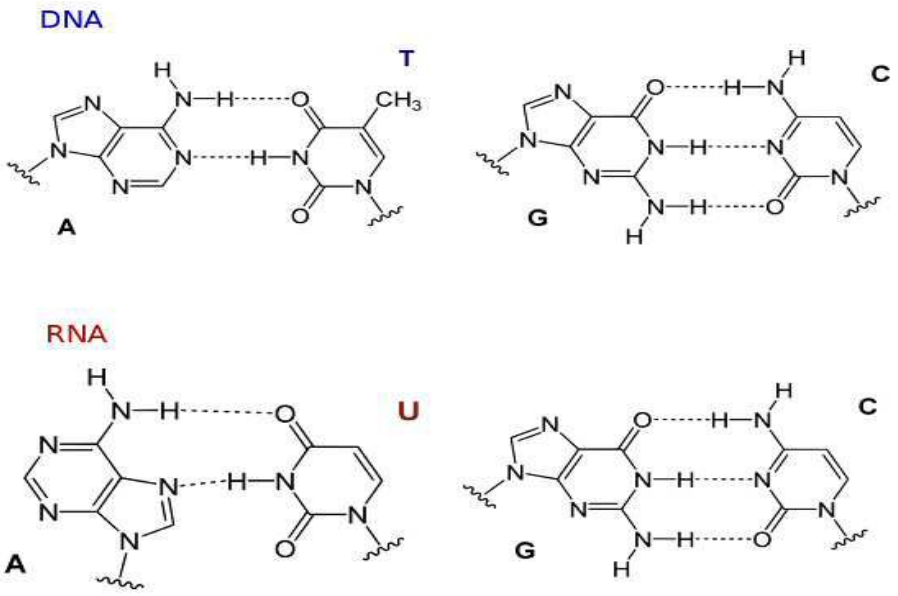
[0091] 실시예는 디제너레이트 염기를 이용하여 이전에 보고된 DNA 기반 데이터 저장 시스템의 순정보 용량과 물리적 밀도를 실험적으로 두 배로 늘릴 수 있음을 보여준다. 또한, 시뮬레이션을 통해 실시예에 기초한 플랫폼의 확장성과 비용 경쟁력을 입증했다. 실시 예는 모든 디제너레이트 염기를 사용할 수 있는 컬럼 기반 올리고 뉴클레오티드 풀 합성 설정(column-based oligonucleotide pool synthesis setup)에 이용될 수 있다. 합성 설정은 뉴클레오타이드 조합의 낮은 편차로 뉴클레오티드의 비율을 정확하게 조절해야 한다. 그런 다음 대규모 실험에서 플

랫폼을 최적화하여, 시뮬레이션에서 제안된 비-등가 비 (non-equivalent ratios)를 이용한 변조된 디제너레이트 염기를 사용할 수 있다.

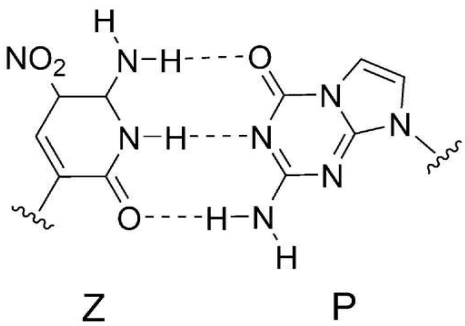
[0092] 또한, 합성 염기에 대한 합성 및 시퀀싱 방법이 개발되면, 변조된 디제너레이트 염기들은 다른 유형의 추가 인코딩문자로 사용될 수 있다. 제1 및 제2 실험 예에서는 디제너레이트 염기가 DNA 염기(A, T, G, C) 중 적어도 두 개의 염기의 혼합으로 구성되었으나, 발명이 이에 한정되는 것은 아니다. 예를 들어, DNA 이외의 RNA 염기나, PNA, XNA, ZNA 등을 포함하는 핵산 유사체(Nucleic acid analogue, https://en.wikipedia.org/wiki/Nucleic_acid_analogue)들은 화학적 합성법이 알려져 있고, 현재 기존에 알려져 있는 효소(Enzyme)나 이의 변형물로 증폭이 가능하다고 알려져 있다. 따라서, 이들에 대한 시퀀싱 방법의 개발에 따라, 다른 유형의 인코딩 문자에 대응하는 디제너레이트 염기로 사용 될 수 있다.

[0093] 국제 공개 특허 문헌 WO 2017/011492 A1 에서는, DNA 이외의 RNA, 그리고 Z, P, dNaM, dSSIC, isoC, isoG, Ds, Px, PNA(Peptide nucleic acid), XNA(Xeno nucleic acid), ZNA(Zip nucleic acid) 등을 포함하는 Nucleic acid analogue 들이 화학적 합성 및 중합효소(polymerase)를 이용한 증폭이 가능한 점이 개시되어 있다. Nucleic acid analogue 는 DNA, RNA와 구조가 유사하지만 backbone이 phosphate backbone이 아니거나 base의 구조가 A, G, T, C, U와 상이한 생화학 물질을 뜻한다.

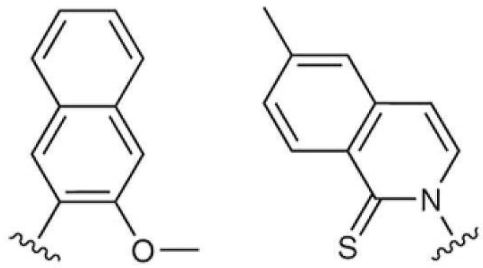
[0094] 실시 예에 따른 디제너레이트 염기는 DNA, RNA, 및 핵산 유사체 중 적어도 두 종류가 혼합된 것으로 정의될 수 있고, 해당 디제너레트 염기에 대응하는 인코딩 문자는 데이터 저장에 이용 될 수 있다. 아래는 DNA 염기, RNA 염기, 및 핵산 유사체들 각각의 분자 구조와 염기 페어링(base paring) 분자 구조이다.



[0095]

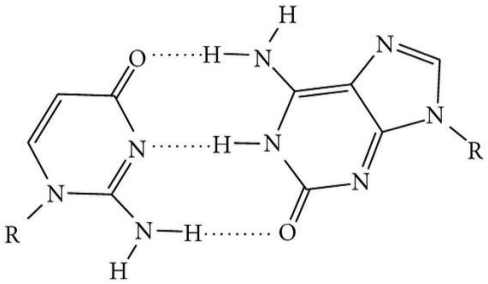


[0096]



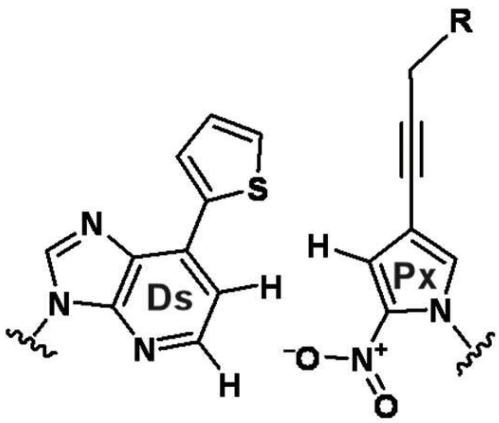
dNaM-d5SICS

[0097]



isoC isoG

[0098]



[0099]

[0100] 이상에서 본 개시에서 설명된 것에 의해 본 발명의 권리범위가 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

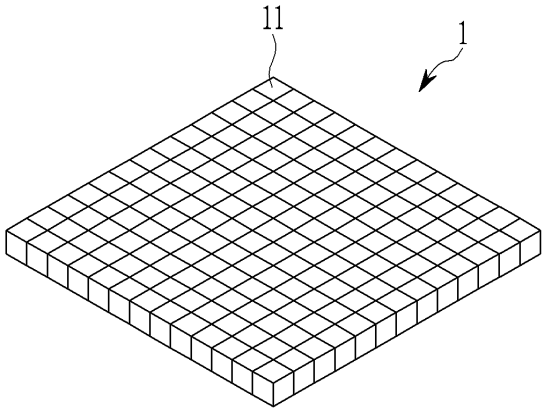
부호의 설명

[0101] 1: DNA 디지털 데이터 저장 장치

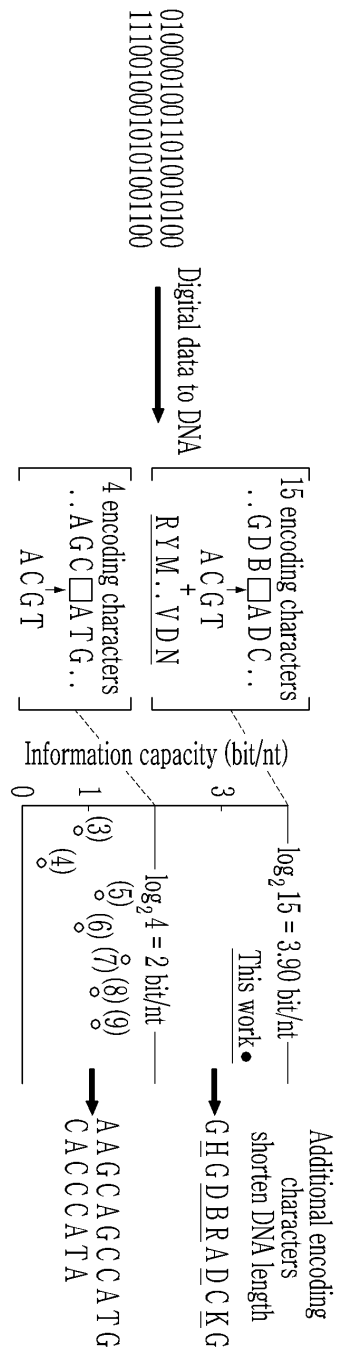
11: 메모리 셀

도면

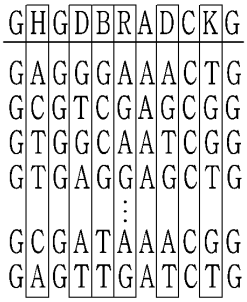
도면1



도면2a


01000010011010010100
11100100001010100100
Digital data to DNA
15 encoding characters
. . G D B □ A D C . .
A C G T
R Y M . . V D N
4 encoding characters
. . A G C □ A T G . .
A C G T
Information capacity (bit/nt)
0
1
3
log2 4 = 2 bit/nt
log2 15 = 3.90 bit/nt
(3)
(4)
(5)
(6)
(7)
(8)
(9)
A A G C A G C C A T G
C A C C C A T A
Additional encoding characters
shorten DNA length
This work
G H G D B R A D C K G

도면2b

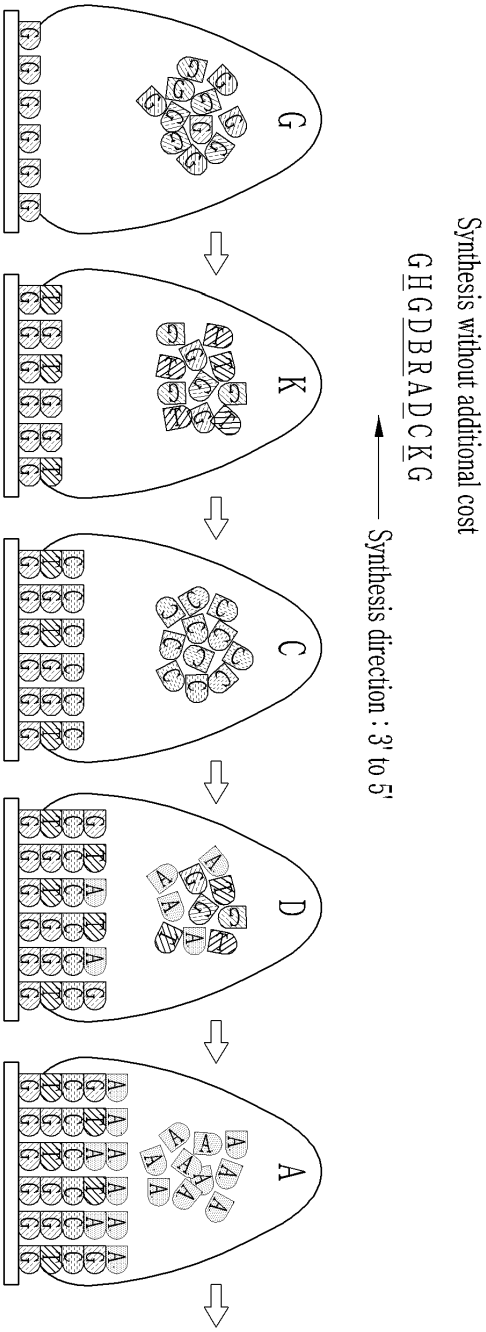


Degenerate bases

Symbol	Base mix
R	A, G
Y	C, T
M	A, C
K	G, T
S	C, G
W	A, T

Symbol	Base mix
H	A, C, T
B	C, G, T
V	A, C, G
D	A, G, T
N	A, C, G, T

도면2c



도면2d

BiNEL.txt — 편집됨

The BiNEL (Biophotonics and Nano Engineering Lab) is located at the Seoul National University. Professor Sunghoon Kwon's group is operated since 2006.

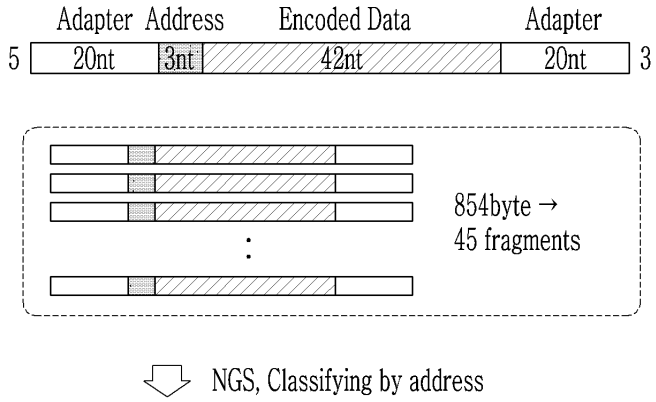
Current members :

Junhoi Kim, Hunjong Na, Sungsik Kim, Dong Yoon Oh, Daewon Lee, Sangwook Bae, Yeongjae Choi, Seowoo Song, Yunjin Jeong, Okju Kim, Seohee Chang, Sudeok Kim, Amos Chungwon Lee, Huiran Yeom, Tae Geun Lim, Hyun Yong Jeong, Jinsung Noh, Jinhyun Kim, SeongKyu Cho, Gi Yoon Lee, Hansol Choi, Yongju Lee, Hyunho Lee, Yonghee Lee, Wonseok Choi, Sumin Lee, Unah Kim, Jinwoo Hyun, HongKeun Oh, Keum Hee Hwang

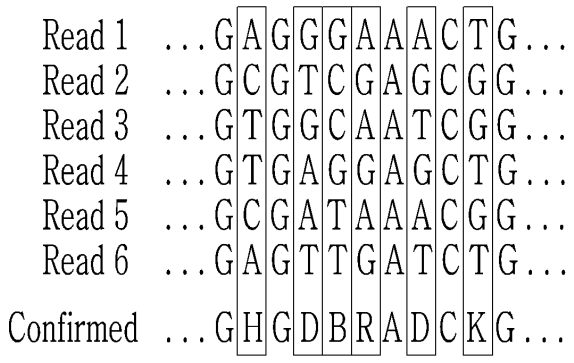
Alumni:

Hyung Jong Bae, Jungmin Kim, Younghoon Song, Yushin Jung, Taehoon Ryu, Jungil Choi, Dongyoung Lee, Sangkwon Han, Howon Lee, Jisung Jang, Jiyun Kim, Jaekyung Koh, Eun Geun Kim, Saifullah Lone, Taehong Kwon, Hyoki Kim, Su Eun Chung, Wook Park, Na Ri Kim, Sung-Eun Choi

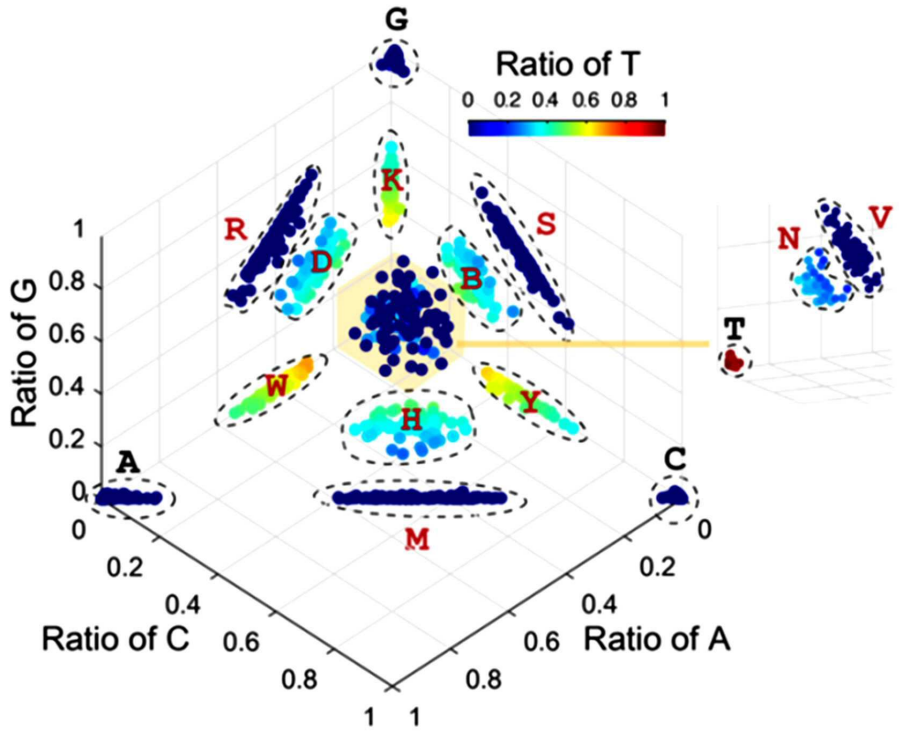
도면3a



도면3b

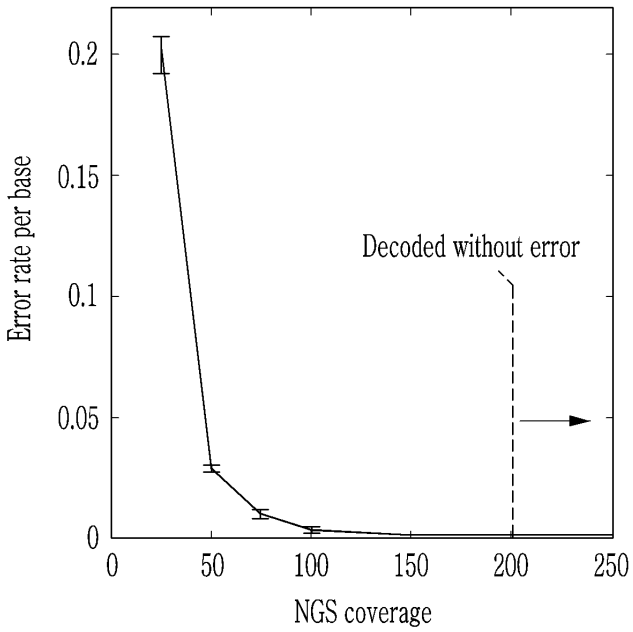


도면3c



Categorizing sequencing data into encoding characters

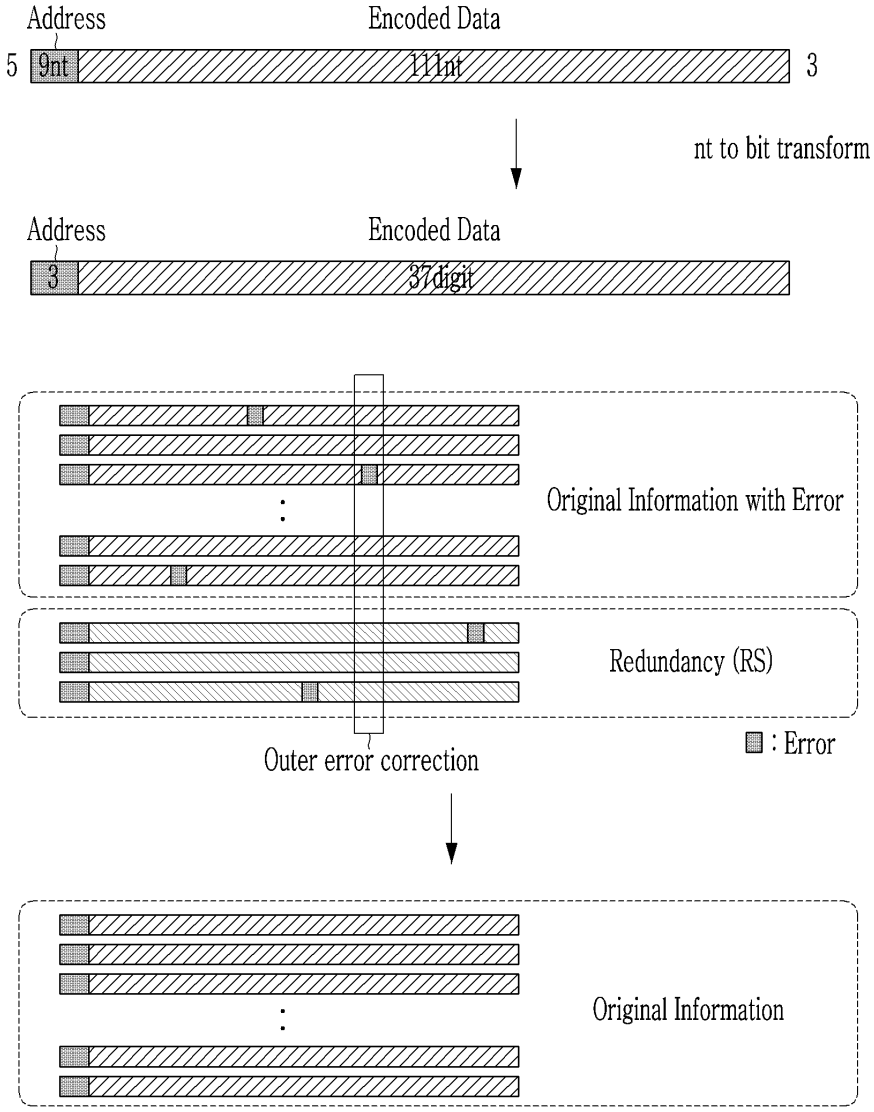
도면3d



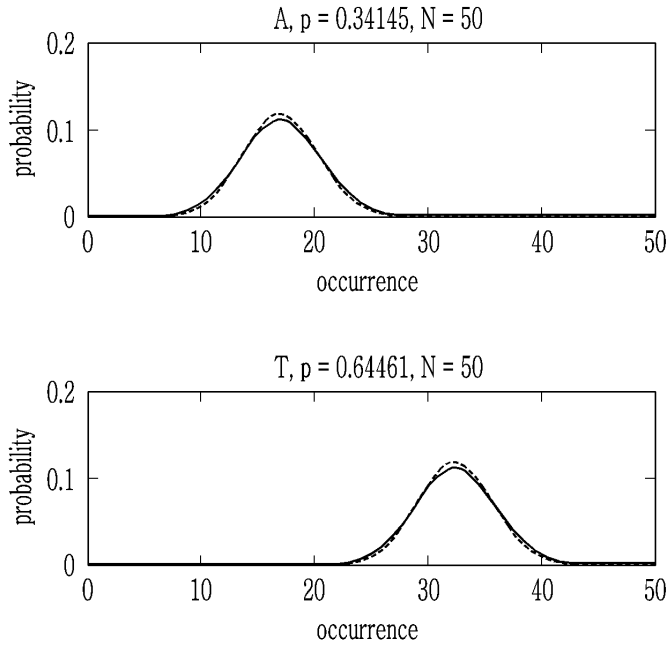
도면4


大東千古閒矇矓
用字例
初聲ㄱ。如감爲柿。골爲蓋。ㄱ。如우
ㅱ爲未舂稻。콩爲大豆。ㅇ。如러울
爲獺。서에爲流澌。ㄷ。如뒤爲茅。담
爲墻。ㅌ。如고티爲蘭。두텁爲蟾蜍。
ㄴ。如노로爲獐。납爲猿。ㅂ。如모롤爲
蠭。ㅍ。如파爲葱。플爲蠅。ㅁ。
如ᄆᆡ爲山。마爲薯蕷。ᄇᆞᆼ。如사ᄇᆡ爲
蝦。드뵈爲貙。ㅈ。如자爲尺。조히爲
紙。ㅊ。如체爲簁。채爲鞭。ㅅ。如손爲
手。셤爲島。ㆆ。如부헝爲鵂鶹。힘爲
筋。ㅇ。如비육爲鷄雛。ᄇᆞ암爲蛇。ㄹ。
如무뤼爲雹。어름爲氷。ㅿ。如아ᄉᆞ
爲弟。너ㅿ爲鴇中聲。ㆍ。如톡爲頓。
ᄭᅩᆺ爲小豆。두리爲橋。ㄱ。래爲楸。一

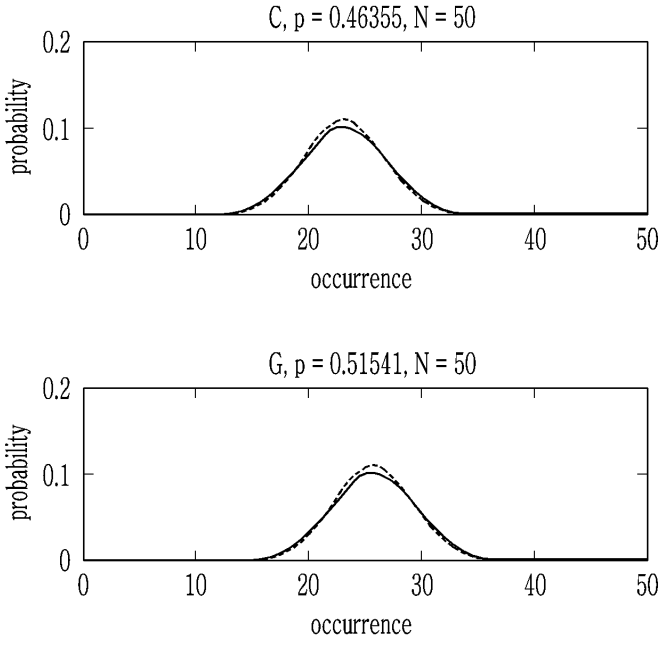
도면5


Address
Encoded Data
5
9nt
111nt
3
nt to bit transform
Address
Encoded Data
3
37digit
Original Information with Error
Redundancy (RS)
: Error
Outer error correction
Original Information

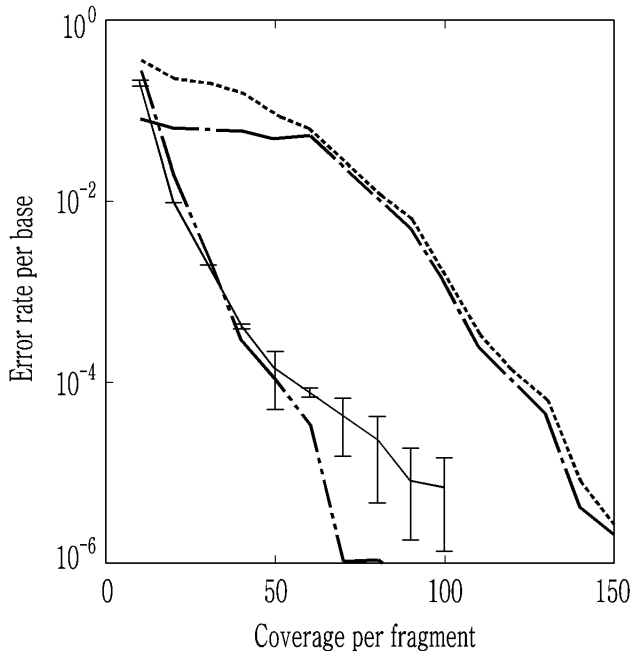
도면6a


A, p = 0.34145, N = 50
probability
0.2
0.1
0
0
10
20
30
40
50
occurrence
T, p = 0.64461, N = 50
probability
0.2
0.1
0
0
10
20
30
40
50
occurrence

도면6b


C, p = 0.46355, N = 50
probability
0.2
0.1
0
0
10
20
30
40
50
occurrence
G, p = 0.51541, N = 50
probability
0.2
0.1
0
0
10
20
30
40
50
occurrence

도면7a



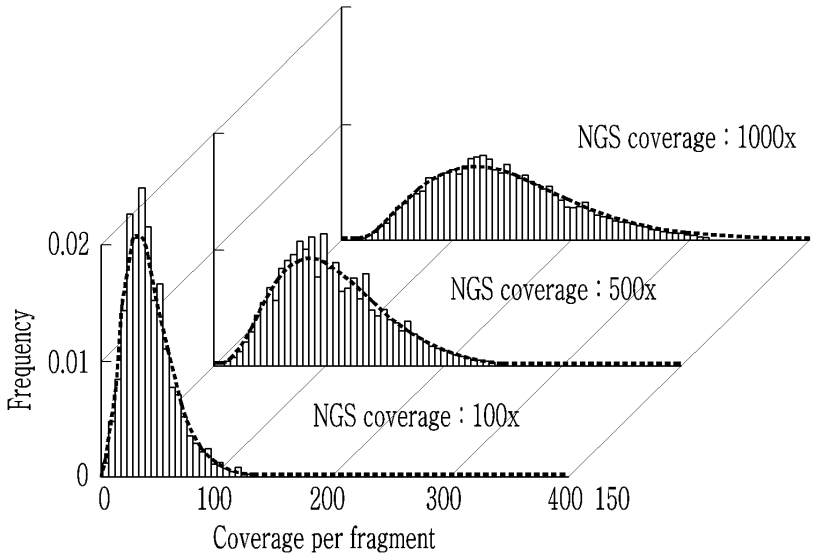
A, C, G, T, W, S(6) - Experiment

A, C, G, T, W, S(6) - Simulation

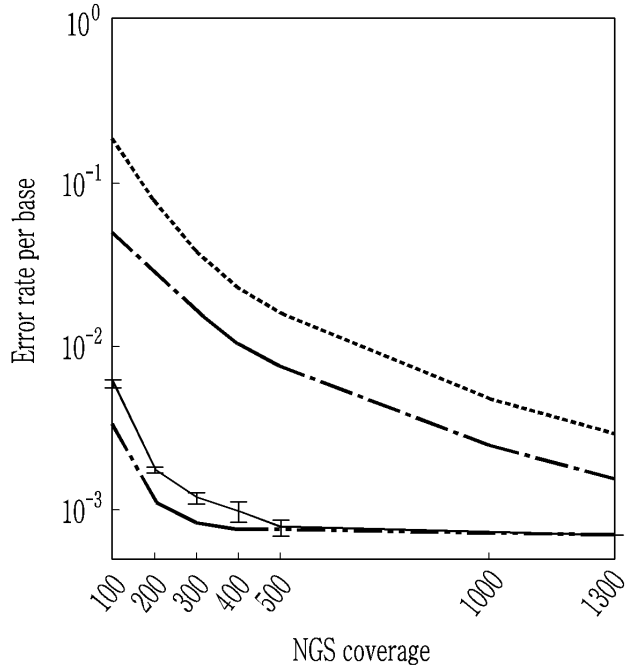
A, C, G, T, R, Y, M, K, S, W, H, V, D, N(15) - Simulation

A, C, G, T, [R, Y, M, K, S, W - ratio of bases mixed 3:7 and 7:3], H, V, D, N(21) - Simulation

도면7b



도면7c



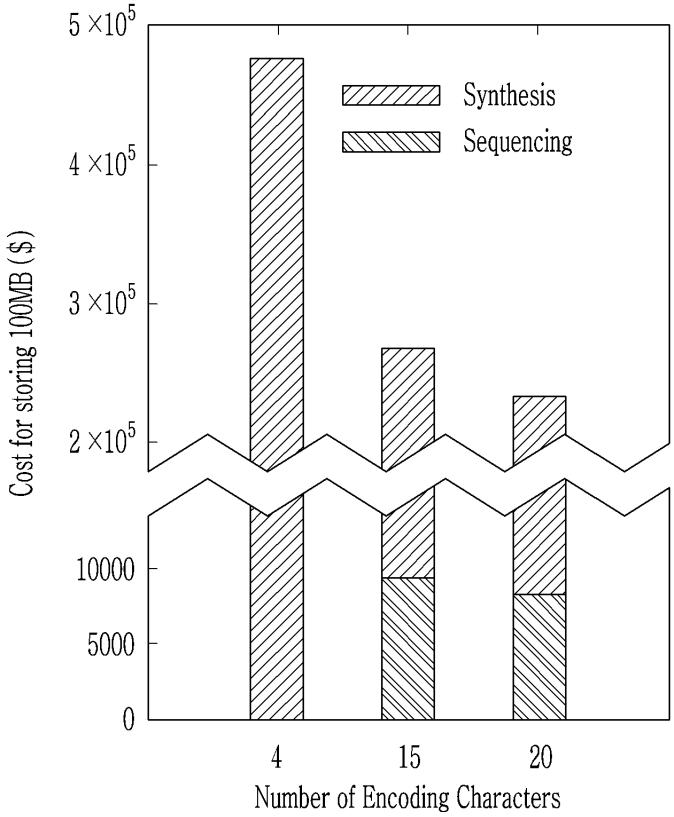
A, C, G, T, W, S(6) - Experiment

A, C, G, T, W, S(6) - Simulation

A, C, G, T, R, Y, M, K, S, W, H, V, D, N(15) - Simulation

A, C, G, T, [R, Y, M, K, S, W - ratio of bases mixed 3:7 and 7:3], H, V, D, N(21) - Simulation

도면8


5 ×10^5
4 ×10^5
3 ×10^5
2 ×10^5
10000
5000
0
Cost for storing 100MB ($)
Synthesis
Sequencing
4
15
20
Number of Encoding Characters