(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0016572
(43) 공개일자 2015년02월12일

(51) 국제특허분류(Int. Cl.)
G06F 19/22 (2011.01) *G06N 3/12* (2006.01)
(21) 출원번호 10-2014-7035408
(22) 출원일자(국제) 2013년05월31일
심사청구일자 없음
(85) 번역문제출일자 2014년12월17일
(86) 국제출원번호 PCT/EP2013/061300
(87) 국제공개번호 WO 2013/178801
국제공개일자 2013년12월05일
(30) 우선권주장
61/654,295 2012년06월01일 미국(US)

(71) 출원인
유럽피안 몰레큘러 바이올로지 래보러토리
독일, 하이델버그 디-69117, 메예르호프스트라세 1
(72) 발명자
골드만, 닉
영국, 케임브리지 케임브리지셔 씨비10 1에스디, 힌스턴, 웰컴 트러스트 게놈 캠퍼스, 유러피안 바이오인포매틱스 인스티튜트
버니, 존
영국, 케임브리지 케임브리지셔 씨비10 1에스디, 힌스턴, 웰컴 트러스트 게놈 캠퍼스, 유러피안 바이오인포매틱스 인스티튜트
(74) 대리인
특허법인씨엔에스

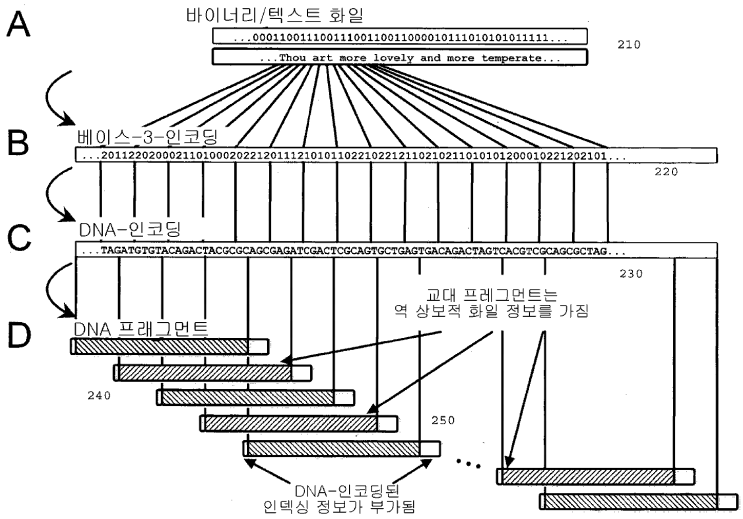
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **DNA 디지털 정보의 고-용량 저장**

(57) 요 약

정보 아이템(210)의 저장 방법이 개시된다. 상기 방법은 바이트(720)를 정보 아이템(210)으로 인코딩하고, 인코딩된 바이트를 스키마를 사용하여 DNA 뉴클레오타이드로 리프리젠테이션하여 DNA 서열(230)을 생성하는 것을 포함한다. 상기 DNA 서열(230)을 복수의 오버래핑 DNA 세그먼트들(240)로 분할하고, 상기 복수의 DNA 세그먼트들(240)에 인덱싱 정보(250)를 부가한다. 마지막으로, 상기 복수의 DNA 세그먼트들(240)을 합성하고(790), 저장한다(795).

대 표 도



특허청구의 범위

청구항 1

- 바이트(bytes)를 정보 아이템(210)으로 인코딩하는 단계;

- 스키마(schema)를 사용하여 인코딩된 바이트를 적어도 하나의 DNA 뉴클레오타이드로 리프리젠테이션하여(720) DNA 서열을 생성하는 단계;

- 상기 DNA 서열을 복수의 오버래핑 DNA 세그먼트들(240)로 분할하는 단계(730);

- 상기 복수의 DNA 세그먼트들(240)에 인덱싱 정보를 부가하는(augmenting) 단계(740);

- 상기 복수의 DNA 세그먼트들(240)을 합성하는 단계(790); 및

- 합성 DNA 세그먼트들(240)을 저장하는 단계

를 포함하는 정보 아이템(210)의 저장 방법.

청구항 2

제1항에 있어서,

상기 합성 DNA 세그먼트들에 어댑터를 부가하는 것을 더욱 포함하는, 정보 아이템의 저장 방법.

청구항 3

제1항 또는 제2항에 있어서,

상기 바이트를 인코딩하기 위해 염기-3 스킴(base-3 scheme)을 사용하는, 정보 아이템의 저장 방법.

청구항 4

제1항 내지 제3항 중 어느 한 항에 있어서,

사용되는 리프리젠테이션 스키마는 DNA 뉴클레오타이드들의 인접한 것들이 다르도록 디자인되는, 정보 아이템의 저장 방법.

청구항 5

제1항 내지 제4항 중 어느 한 항에 있어서,

인덱싱 정보에 패리티 검사(parity-check)를 추가하는 단계를 더욱 포함하는, 정보 아이템의 저장 방법.

청구항 6

제1항 내지 제5항 중 어느 한 항에 있어서,

상기 합성 DNA 세그먼트들(240)의 교대하는(alternate) 것들은 역 상보적인, 정보 아이템의 저장 방법.

청구항 7

제1항 내지 제6항 중 어느 한 항에 있어서,

사용되는 리프리젠테이션 스키마는 긴, 자가-역 상보적 합성 DNA 세그먼트들을 회피하도록 디자인되는, 정보 아이템의 저장 방법.

청구항 8

제1항 내지 제7항 중 어느 한 항에 있어서,

합성 DNA 세그먼트들을 디코딩하여 상기 정보 아이템을 재구성하는 단계를 더욱 포함하는, 정보 아이템의 저장 방법.

청구항 9

제1항 내지 제8항 중 어느 한 항의 방법에 따라 저장된 정보 아이템을 포함하는, 비휘발성, 비일시적 저장 매체.

청구항 10

제1항 내지 제8항 중 어느 한 항에 따른 방법을 실행하는 로직(logic)을 포함하는 컴퓨터 프로그램 제품.

명 세 서

기 술 분 야

[0001] 본 발명은 DNA 디지털 정보의 저장을 위한 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] DNA는 장기간 동안 콤팩트 형태로 쉽게 저장되는 방대한 양의 정보를 보유하기 위한 능력을 갖는다[1,2]. 디지털 정보를 위한 저장으로서 DNA를 이용한 아이디어는 1995년 이래로 존재하였다[3]. DNA 저장의 물리적 구현은 지금까지 사소한 양 - 전형적으로 영문의 몇 개의 숫자 또는 단언 - 의 정보만을 저장하였다[4-8]. 발명자들은 자기 물질 또는 광학 물질상의 데이터 저장 대신에, 물리적인 DNA에서의 인코딩된 임의의 크기의 디지털 정보의 대규모 저장 및 복구에 대해 인식하지 못하고 있었다.

[0003] 현재 DNA의 합성은 생의학 적용에 중점을 둔 전문화된 기술이다. DNA 합성 비용은 지난 십년에 걸쳐 꾸준히 감소되었다. 본 명세서에 기재된 바와 같이, DNA 분자에서의 시간 척도 데이터 저장은 3 내지 5년 마다 새로운 매체로 드물지만 규칙적으로 이송하는 테이프상에서 데이터 저장에 대한 현재의 장기간의 기록보관 프로세스에 비해 DNA 분자에서의 데이터 저장이 어느 기간에 더 비용 효율적일 수 있는지 추측하는 것이 흥미롭다. DNA 합성을 위한 현재의 "규격품(off the shelf)" 기술은 미국 달러 당 약 100바이트의 가격에 해당한다. Agilent Technologies(Santa Clara, CA)로부터 상업적으로 구입가능한 새로운 기술은 이러한 비용을 실질적으로 감소시킬 수 있다. 그러나, 또한 테이프 매체 사이의 데이터의 정기적인 이송이 고려될 필요가 있다. 문제는 이러한 데이터 이송을 위한 비용 및 이러한 비용이 시간 경과에 따라 고정되느냐 감소하느냐 하는 것이다. 만일 실질적인 양의 비용이 고정될 것으로 추정되면, 데이터 저장을 위한 DNA 분자의 사용이 테이프 매체상의 정기적인 데이터 저장에 비해 더 비용 효율적인 시계(time horizon)가 존재한다. 400년(적어도 80번의 매체 이송) 후에, DNA 분자를 이용한 이러한 데이터 저장은 이미 비용 효율적인 것이 가능하다.

[0004] 기존에 취급했던 것 보다 더 많은 정보를 저장하는 실용적인 인코딩-디코딩 방법이 본 명세서에 기재된다. 발

명자들은 5개의 컴퓨터 파일 - 총 757051바이트(739kB)의 하드 디스크 스토리지 및 5.2 x 10^6 비트의 추정 샤논 정보[9] - 을 DNA 코드로 인코딩하였다. 발명자들은 후속적으로 이 DNA를 합성하고, 합성된 DNA를 미국으로부터 영국을 통해 독일로 수송하고, 그 DNA를 시퀀싱하고, 100% 정확도로 5개 컴퓨터 파일들 모두를 재구성하였다.

[0005] 상기 5개의 컴퓨터 파일들은 영문(모두 셰익스피어의 소네트 154편), 고전 과학 논문의 PDF 문서[10], JPEG 칼라 사진 및 26초의 연설(마틴 루서 킹의 "나에게는 꿈이 있습니다(I have A Dream)" 연설)이 담긴 MP3 포맷 오디오 파일을 포함하였다. 이 데이터 저장은 알려진 기존의 DNA-기반의 저장의 약 800배의 정보를 나타내며 상당히 더 많은 다양한 디지털 포맷을 포함한다. 그 결과는, DNA 저장이 점진적으로 현실적이며, 미래에 디지털 정보를 보관하는 비용-효율적인 수단을 제공할 수 있으며, 저 접근성, 수 십년 기록 보관 작업을 위해 이미 비용 효율적일 수 있음을 입증한다.

[0006] 선행 기술

[0007] 용이하게 성취된 조건하에서 정보를 안정하게 저장하기 위한 고 용량 DNA[1,2]는 1995년 이래로 DNA를 정보 저장을 위한 매력적인 표적이 되게 하였다[3]. 정보 밀도에 부가적으로, DNA 분자는 정보 캐리어로서 입증된 실적을 가지며, DNA 분자의 장수성은 알려져 있으며, 지구상의 생명의 기초로서, DNA 분자의 조작, 저장 및 디코딩 방법은 부단한 기술적 혁신의 대상으로 남을 것이며, DNA-기반의 지적 생명체가 유지된다는 사실이 알려져 있다[1,2]. 리빙 벡터 DNA[5-8](생체내 DNA 분자) 및 합성 DNA[4,1](시험관내 DNA) 모두에 기초한 데이터 저장 시스템이 제안된 바 있다. 생체내 데이터 저장 시스템은 여러 결점을 갖는다. 이러한 결점은 리빙 벡터 유기체내의 DNA 분자의 생존능력에 영향을 주지않고 조작될 수 있는 양, 게놈 엘리먼트 및 위치의 제약을 포함한다. 이러한 리빙 벡터 유기체의 예는 이에 한정하는 것은 아니나 박테리아를 포함한다. 생존능력의 감소는 용량 감소 및 정보 인코딩 스킴의 복잡성 증가를 포함한다. 더욱이, 생식 계열 및 체세포 돌연변이는 저장된 정보 및 디코딩된 정보가 시간 경과에 따라 적합도가 감소되는 것을 초래하고, 또한 아마도 리빙 DNA의 저장 조건에 대한 요건이 조심스럽게 조절되어야 할 것이다.

[0008] 이와 상반적으로, "분리된 DNA(isolated DNA)"(즉, 시험관내 DNA)는 보다 용이하게 "기록(written)"되고 수 만 년령이 된 시료로부터 논-리빙 DNA의 예의 통상적인 복구[11-14]은 잘 제조된 논-리빙 DNA 시료가 쉽게-성취되는 저-유지 환경에서(즉, 춥고, 건조하고 어두운 환경)에서 예외적으로 긴 수명을 갖는다는 것을 나타내어 준다[15-17].

[0009] DNA에서 정보(데이터라고도 칭함) 저장에 대한 이전 연구는 전형적으로, 인간이 판독가능한 메시지를 인코딩된 형태의 DNA로 "기록(writing)"하고, 그 다음 그 DNA 서열을 검출하고 그 서열을 디코딩함으로써 인코딩된 인간이 판독가능한 메시지를 "판독(reading)"하는데 중점을 둔다. DNA 컴퓨팅 분야에서 연구는 기본적으로 대규모의 결합된(콘텐츠가 어드레싱된) 메모리를 허용하는 스킴을 제공하여 왔으나[3,18-20], 실용적인 DNA-저장 스킴으로서 이러한 연구를 발달시키기 위한 시도는 없었다. 도 1은 14개의 선행 연구에서 성공적으로 인코딩되고 복구된 정보의 양을 나타낸다(y-축상의 로그 스케일 주목). 14개의 선행 실험(속인 빈 원) 및 본 발명(속인 찬 원)에 대해 포인트를 나타내었다. 이러한 방식으로 저장된 가장 많은 양의 인간이 판독가능한 메시지는 1280 캐릭터의 영어 텍스트[8]이며, 이는 약 6500비트의 샤논 정보[9]에 해당하는 것이다.

[0010] 인도의 과학 및 산업 연구 협의회는 DNA에 정보를 저장하는 방법을 교시한 미국 특허 출원공개 제 US 2005/0053968호(Bharadwaj 등)를 출원하였다. 미국 특허 출원공개 제 US 2005/0053968호의 방법은 확장된 ASCII 캐릭터 셋트의 각 캐릭터를 나타내는 4-DNA 염기들을 사용하는 인코딩 방법을 사용하는 것을 포함한다. 그 다음, 디지털 정보인 암호 키를 포함하는 합성 DNA 분자가 생성되고, 프라이머 서열 각 옆에 플랭킹된다.

최종적으로, 합성된 DNA는 저장 DNA에 편입된다. DNA의 양이 너무 많을 경우에, 그 정보는 다수의 세그먼트로 프래그먼트화될 수 있다. 미국 특허 출원공개 제 US 2005/0053968호에 기재된 방법은 세그먼트 중 하나의 헤더 프라이머를 상기 세그먼트의 후속적인 것에 있는 테일 프라이머와 매칭시킴으로써 프래그먼트화된 DNA 세그먼트를 재구성할 수 있다.

[0011] 정보를 DNA에 저장하는 기술이 기재된 다른 특허 공개문헌이 알려져 있다. 예를 들어, 미국 특허 제 6,312,911호에는 코딩된 메시지를 DNA내에 감추는 스테가노그래피 방법이 기재되어 있다. 상기 방법은 게놈 DNA 인코딩된 메시지를 게놈 DNA 시료내에 은폐한 다음, 상기 DNA 시료를 마이크로도트로 더욱 은폐하는 것을 포함한다. 미국 특허 제 6,312,911호는 특히 기밀 정보의 은폐를 위한 것이다. 이러한 정보는 일반적으로 제한된 길이를 가지며, 이에 따라 그 문서는 보다 긴 길이를 갖는 정보의 아이템을 저장하는 방법에 대해 언급하지 않는다. 이와 동일한 발명자는 국제 공개 제 WO 03/025123호로서 공개된 국제 특허출원을 출원하였다.

발명의 내용

해결하려는 과제

[0012] 정보 아이템을 저장하는 방법이 기재된다. 상기 방법은 바이트(bytes)를 정보 아이템으로 인코딩하는 것을 포함한다. 인코딩된 바이트는 인-실리코(in-silico)에서 DNA 서열을 생성하는 스키마(schema)를 이용하여 DNA 뉴클레오타이드로 리프리젠테이션된다. 다음 단계에서, 그 DNA 서열은 복수의 오버래핑 DNA 세그먼트로 분할되고, 인덱싱 정보가 그 복수의 DNA 세그먼트에 부가된다. 최종적으로, 복수의 DNA 세그먼트가 합성 및 저장된다.

[0013] 인덱싱 정보를 DNA 세그먼트에 부가하는 것은, 정보 아이템을 나타내는 DNA 서열내의 세그먼트의 위치가 독특하게 확인될 수 있음을 의미한다. 헤드 프라이머와 테일 프라이머를 매칭하는 것에 의존할 필요가 없다. 이는 세그먼트들 중 하나가 올바르게 재생되는 것이 실패하더라도 거의 전체 정보 아이템을 복구하는 것이 가능하다. 인덱싱 정보가 존재하지 않을 경우에, DNA 서열 내의 위치가 명확히 확인될 수 없는 "오펀(orphan)" 세그먼트에 기인하여 세그먼트들이 서로 매칭될 수 없다면, 전체 정보 아이템을 올바르게 재생하는 것이 불가능할지도 모르는 위험이 있다.

[0014] 오버래핑 DNA 세그먼트의 사용은 중복도가 정보 아이템의 스토리지에 형성되는 것을 의미한다. DNA 세그먼트들 중 하나가 디코딩될 수 없는 경우에, 인코딩된 바이트는 그 DNA 세그먼트들 중 이웃한 것으로부터 여전히 복구될 수 있다.

[0015] DNA 세그먼트의 다중 카피는 알려진 DNA 합성 기술을 이용하여 제조될 수 있다. 이는 DNA 세그먼트의 카피의 일부가 변질되고 디코딩될 수 없을지라도 정보 아이템의 디코딩을 가능케 하는 부가적인 중복도를 제공한다.

과제의 해결 수단

[0016] 본 발명의 일 견지로, 인코딩하는데 사용되는 대표적인 스키마는, DNA 뉴클레오타이드들의 인접한 것들이 다르도록 디자인된다. 이는 DNA 세그먼트의 합성, 재생 및 시퀀싱(판독)의 신뢰도를 증가시키기 위한 것이다.

[0017] 본 발명의 다른 견지로, 패리티 검사(parity-check)가 인덱싱 정보에 부가된다. 이 패리티 검사는 DNA 세그먼트의 잘못된 합성, 재생 또는 시퀀싱을 확인해 줄 수 있다. 패리티 검사는 또한 오류 수정 정보를 포함하도록 확장될 수 있다.

[0018] 합성 DNA 세그먼트들과 교대하는(alternate) 것들은 역 상보적이다. 이들은 DNA에서 부가적인 중복도를 제공하며, DNA 세그먼트의 어느 것이 변질된 경우에 이용가능한 보다 많은 정보가 존재함을 의미한다.

도면의 간단한 설명

[0019] 도 1은 DNA에 저장된 정보 및 성공적으로 복구된 정보의 양을 시간에 따라 나타낸 그래프이다.

도 2는 본 발명의 방법의 일 예를 나타낸 것이다.

도 3은 저장의 비용 효율성을 기간 경과에 걸쳐 나타낸 그래프를 보여준다.

도 4는 자가-역 상보적 패턴을 갖는 모티프를 나타낸 것이다.

도 5는 인코딩 효율성을 나타낸 것이다.

도 6은 오류율을 나타낸 것이다.

도 7은 상기 방법의 인코딩 흐름도를 나타낸 것이다.

도 8은 상기 방법의 디코딩 흐름도를 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0020] 지금까지 DNA 저장의 실용적인 실행을 위한 주요 도전 중 하나는 특정 디자인에 대하여 긴 DNA 서열을 생성하는 것에 대한 어려움이었다. 긴 DNA 서열은 긴 텍스트 아이템 및 비디오와 같은 큰 데이터 파일을 저장하는데 요구된다. 각 디자인된 DNA의 복수의 카피로 인코딩을 이용하는 것이 또한 바람직하다. 이러한 중복(redundancy)은 하기에 설명되는 바와 같이 인코딩 오류 및 디코딩 오류 모두에 대해 보호한다. 각각의 (잠재적으로 큰) 메세지를 인코딩하기 위해 개별적인 긴 DNA 사슬에 기초한 시스템을 사용하는 것은 비용 효율적이지 못하다[8]. 발명자들은 DNA 세그먼트들의 각각과 관련된 '인덱싱' 정보를 사용하여 전체 메세지를 인코딩하는 가상의 긴 DNA 분자에서 DNA 세그먼트의 위치를 나타내는 방법을 개발하였다.

[0021] 발명자들은 기존의 고 처리량 기술에서 높은 오류율과 관련된 것으로 알려진 DNA 호모폴리머(즉, 하나 이상의 동일한 염기의 런(run))를 금지시키는 것을 포함하는 DNA 세그먼트로부터 인코딩된 메세지의 복구 가능성을 증가시키는 코드 이론의 방법을 사용하였다. 또한, 발명자들은 패리티 검사 비트(parity-check bit)[9]와 유사한 단순한 오류-검출 요소를 코드내의 인덱싱 정보내로 포함시켰다. 이에 한정하는 것은 아니나, 오류-수정 코드[9] 및 실제로, 정보학에서 현재 사용되는 실질적으로 어느 형태의 디지털 데이터 보안(예, RAID-기반의 스킴[21])을 포함하는 보다 복잡한 스킴이 미래의 DNA 저장 스킴[3]의 개발에 실행될 수 있다.

[0022] 발명자들은 본 발명의 DNA 저장용 개념 증명으로서 인코딩될 5개의 컴퓨터 파일을 선택하였다. 상기 파일을 인간-판독가능한 정보로 한정하기 보다는, 다양한 일반적인 포맷을 이용한 파일들을 선택하였다. 이는 임의적인 타입의 디지털 정보를 저장하는 본 발명의 교시의 능력을 입증하였다. 파일들은 모두 154개의 세익스피어의 소네트(TXT 포맷), ref. 10의 전체 텍스트 및 도면(PDF 포맷), EMBL-European Bioinformatics Institute의 중간-해상도 칼라 사진(JPEG 2000 포맷), 마틴 루서 킹의 "나에게는 꿈이 있습니다(I have A Dream)" 연설의 26초의 발췌(MP3 포맷) 및 바이트를 염기-3 디지트(base-3 digits)로 변환하기 위해 본 연구에 사용된 허프먼(Huffman) 코드를 정의하는 파일(인간-판독가능한 텍스트 파일로서)을 포함하였다.

[0023] DNA-저장을 위해 선택된 5개의 파일들은 다음과 같다:

[0024] wssnt10.txt -- 107738 bytes -- ASCII 텍스트 포맷, 모두 154 세익스피어 소네트(Project Gutenberg,

http://www.gutenberg.org/ebooks/1041로부터)

[0025] watsoncrick.pdf -- 280864 bytes -- PDF 포맷 문서, DNA의 구조를 설명하는 Watson 및 Crick(1953)의 간행물[10](*Nature* website, http://www.nature.com/nature/dna50/archive.html로부터, 고압축하여 보다 작은 파일 크기로 성취하도록 변형됨).

[0026] EBI.jp2 -- 184264 bytes -- EMBL-European Bioinformatics Institute의 JPEG 2000 포맷 이미지 파일 칼라 사진(16.7M 칼라, 640 x 480 픽셀 해상도)(보유 사진).

[0027] MLK_excerpt_VBR_45-85.mp3 -- 168539 bytes -- MP3 포맷 소리 파일, 마틴 루서 킹의 "나에게는 꿈이 있습니다(I have A Dream)" 연설의 26초의 발췌(http://www.americanrhetoric.com/speeches/mlkihaveadream.htm로부터, 고압축을 성취하도록 변형됨, 가변 비트 전송률, 전형적으로 48-56kbps; 샘플링 주파수 44.1kHz)

[0028] View_huff3.cd.new -- 15646 bytes -- ASCII 파일, 바이트를 염기-3 디지트(트리츠(trits))로 변환하기 위해 본 시험에 사용된 허프먼 코드를 정의하는 인간-판독가능한 파일

[0029] 상기 5개의 컴퓨터 파일은 총 757051 bytes를 포함하며, 이는 5.2×10^6 비트의 샤논 정보와 대략적으로 동등하거나 또는 저장된 것으로 알려진 기존의 최대량과 같이 인코딩되고 복구된 인간이 설계한 정보의 800배에 해당한다.

[0030] 각각의 컴퓨터 파일을 인코딩하는 DNA를 소프트웨어를 이용하여 산출하였으며, 그 방법을 도 7에 나타내었다. 본 명세서에 나타낸 본 발명의 일 견지(700)로, 각 컴퓨터 파일(210)을 포함하는 바이트를, DNA 서열(230)을 형성하는 5 또는 6 염기(하기 참조)에 의해 각 바이트가 대체되는 인코딩된 파일(220)을 생성하기 위한 인코딩 스킴에 의해 호모폴리머를 갖지 않은 DNA 서열(230)로서, 단계 720에 나타내었다. 인코딩 스킴에 사용된 코드는 런 길이-제한 채널을 위해 최적의 정보 용량에 근접한 단순 인코딩을 허용하도록 구성되었다(즉, 반복 뉴클레오타이드를 함유하지 않음). 그러나, 다른 인코딩 스킴이 사용될 수 있는 것이 인식될 것이다.

[0031] 그 결과 형성된 인 실리코(*in silico*) DNA 서열(230)은 표준 올리고뉴클레오타이드 합성에 의해 용이하게 생성되기에는 너무 길다. 따라서, 각각의 DNA 서열(230)은 75염기의 오버랩을 갖는 길이 100염기의 오버래핑 세그먼트들(240)로 단계 730에서 분할되었다. 어느 특정 염기 런에 도입된 체계적인 합성 오류의 위험을 줄이기 위해, 상기 세그먼트들와 교대하는 것들은 단계 740에서 이들의 역 상보물로 변환되었으며, 이는 각 염기가 각 방향으로 두 번, 4회 "기록"됨을 의미한다. 그 다음, 각 세그먼트는 세그먼트(240)가 유래한 컴퓨터 파일 및 그 컴퓨터 파일(210)내의 이의 위치, 그리고 단순 오류-검출 정보의 결정을 가능케 하는, 인덱싱 정보(250)로 단계 750에서 증강되었다(augmented). 이 인덱싱 정보(250)는 또한 비-반복 DNA 뉴클레오타이드로서 단계 760에서 인코딩되었으며, 단계 770에서 DNA 세그먼트(240)의 100 정보 저장 염기에 첨부되었다. 75 염기의 오버랩을 갖는 100 염기의 길이로의 DNA 세그먼트(240)의 분할이 순전히 임의적(ar비트rary)이라는 것이 인식될 것이다. 다른 길이 및 오버랩이 사용될 수 있으며, 이로써 본 발명을 제한하는 것은 아니다.

[0032] 통틀어, 모두 5개의 컴퓨터 파일은 153335 스트링의 DNA로 나타내어졌다. 각각의 DNA 스트링은 117 뉴클레오타이드를 포함하였다(본래의 디지털 정보 + 인덱싱 정보를 인코딩함). 사용된 인코딩 스킴은 합성 DNA의 다양한 특징(예, 균일한 세그먼트 길이, 호모폴리머의 부재)을 가졌으며, 이는 그 합성 DNA가 천연(생물학적) 오리진을 갖지 않았다는 것을 명백히 하였다. 따라서, 그 합성 DNA는 의도적인 디자인 및 인코딩된 정보를 갖는다는 것

이 명백하다[2].

[0033] 상술한 바와 같이, DNA 세그먼트(240)에 대한 다른 인코딩 스킴은, 예를 들어, 향상된 오류-수정 특성을 제공하기 위해 사용될 수 있다. 이는 또한, 더 많거나 더 큰 파일들이 인코딩될 수 있도록 인덱싱 정보의 양을 증가시키기 위해 단순(straightforward)할 수 있다. 네스티드 프라이머 분자 메모리(Nested Primer Molecular Memory, NPMM) 스킴[19]이 16.8M 고유 어드레스[20]에서 이의 실용적인 최대 용량에 도달하는 것이 제시되었으며, 본 발명의 방법이 거의 임의로 다량의 정보를 인코딩할 수 있도록 하기 위해 그 이상으로 확장될 수 없는 이유는 없는 것으로 보인다.

[0034] DNA 세그먼트(240)에서 체계적인 패턴을 회피하기 위한 코딩 스킴에 대한 한 확장이 정보에 변화를 부가하는 것일 수 있다. 이를 수행하기 위한 두 가지 방법이 시도되었다. 첫 번째 방법은 DNA 세그먼트(240)내에 정보의 "셔플링(shuffling)"을 포함하였다. 그 정보는 셔플링의 패턴을 알게되면 복구될 수 있다. 본 발명의 일 견지로, 다양한 패턴의 셔플이 상이한 DNA 세그먼트들에 대해 사용되었다.

[0035] 다른 방법은 각각의 DNA 세그먼트(240)내의 정보에 소정의 임의성(randomness)을 부가하는 것이다. 일련의 랜덤 디지트(random digit)가 이를 위해 사용될 수 있으며, 이는 일련의 랜덤 디지트의 모듈식 첨가를 이용하며, 상기 디지트는 DNA 세그먼트(240)내에 인코딩된 정보를 포함한다. 그 정보는 사용된 일련의 랜덤 디지트를 알게되면 디코딩 중에 모듈식 감산에 의해 쉽게 복구될 수 있다. 본 발명의 일 견지로, 다양한 DNA 세그먼트(240)에 대해 다양한 일련의 랜덤 디지트가 사용되었다.

[0036] 단계 720에서 디지털 정보 인코딩이 다음과 같이 수행되었다. 하드-디스크 드라이브에 저장된 (도 2a에 나타낸) 디지털 정보의 5개 컴퓨터 파일(210)을 소프트웨어를 이용하여 인코딩하였다. 단계 720에서 인코딩될 각각의 5개 컴퓨터 파일(210)의 각 바이트는 (하기의) 표 1에 열거된 목적대로 설계된 허프먼 코드를 이용하여 염기-3 디지트('트리츠(trits)' 0, 1 및 2)를 통해 DNA 염기의 서열로 나타내어 인코딩된 파일(220)을 생성하였다. 이 예시적인 코딩 스킴을 도 2b에 개략적으로 나타내었다. 가능한 256 바이트의 각각은 5 또는 6 트리츠로 나타내어졌다. 후속적으로, 개개의 트리츠는 이전 뉴클레오타이드와 다른 3 뉴클레오타이드로부터 선택된 DNA 뉴클레오타이드로서 인코딩되었다(도 2c). 즉, 본 발명의 이러한 견지를 위해 선택된 코딩 스킴에서, 개개의 3 뉴클레오타이드는 호모폴리머가 없는 것을 확실히 하기 위해 사용된 이전 것과 상이하였다. 그 결과 형성된 DNA 서열(230)은 단계 730에서 도 2d에 나타낸 바와 같이, 길이 100 염기의 DNA 세그먼트(240)로 분할되었다. 개개의 DNA 세그먼트들은 75염기가 이전 DNA 세그먼트에 오버래핑되어 용이하게 합성된 길이의 DNA 세그먼트를 제공하고, 중복성(redundancy)을 제공하였다. 그 DNA 세그먼트와 교대하는 것은 역 상보적이었다.

[0037] 인덱싱 정보(250)는 파일 식별을 위한 2개의 트리츠(본 구현에서는, $3^2 = 9$ 파일이 구별가능하도록 함), 파일내 위치 정보를 위한 12개의 트리츠(파일당 $3^{12} = 531441$ 위치를 확인할 수 있음) 및 하나의 '패리티-검사(parity-check)' 트리트를 포함하였다. 인덱싱 정보(250)는 단계 760에서 비-반복 DNA 뉴클레오타이드로 인코딩되고, 단계 770에서 상기 100 정보 스토리지 염기에 첨부되었다. 각 인덱싱된 DNA 세그먼트(240)는 단계 780에서 '노 호모폴리머(no homopolymers)' 규칙에 부합하는, 각 말단에 추가의 1개의 염기를 가졌으며, 이는 전체 DNA 세그먼트(240)가 본 실험의 '판독(reading)' 단계 동안에 역 상보적이었는지를 나타낸다.

[0038] 통틀어, 5개의 컴퓨터 파일(210)은 153335 스트링의 DNA로 나타내어졌으며, 각각 117(1 + 100 + 2 + 12 + 1 + 1) 뉴클레오타이드(본래의 디지털 정보 및 인덱싱 정보를 암호함)를 포함하였다.

[0039] 본 명세서에 기재된 본 발명의 견지에서 각 스트링의 데이터-암호 요소는 DNA 염기당 5.07 비트에서 샤논 정보를 함유할 수 있으며, 이는 하나로 제한된 런 길이를 갖는 염기-4 채널에 대해 DNA 염기당 5.05 비트의 이론적 최적값에 가까운 것이다. 인덱싱 구현(250)은 3^{14} = 4782969 고유 데이터 위치를 허용한다. 파일 및 파일내 위치를 특정하기 위해 사용된 인덱싱 트리츠 (및 이에 따른 염기)의 수를 단지 2만큼 6으로 증가시키는 것은, NPMM 스킴[19,20]에 대한 실현 가능한 최대값인 16.8M을 초과하여 3^{16} = 43046721 고유 위치를 제공한다.

[0040] 또한, 단계 790의 DNA 합성 공정은 Illumina 시퀀싱 플랫폼에서의 시퀀싱을 촉진하기 위해 각각의 올리고뉴클레오타이드(올리고)의 각 말단에 33bp 어댑터를 편입시키는데 사용되었다.

[0041] 5' 어댑터: ACACTCTTTCCCTACACGACGCTCTTCCGATCT

[0042] 3' 어댑터: AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

[0043] 상기 153335 DNA 세그먼트 디자인(240)은 단계 790에서 종래 기재된 Agilent Technologies의 OLS(올리고 라이브러리 합성(Oligo Library Synthesis)) 공정의 업데이트 버전[22,23]을 사용하여 3개의 구별되는 런으로 합성되어 (상기 DNA 세그먼트(240)는 무작위로 런에 할당됨), 각 DNA 세그먼트 디자인의 약 $1.2x10^7$ 카피를 생성하였다. 오류는 500 염기당 그리고 독립적으로 상기 DNA 세그먼트(240)의 다른 카피에서 단지 약 하나의 오류가 발생하는 것으로 나타났다. Agilent Technologies는 종래에 개발된 포스포아미디트 화학[24]을 적용하고 Agilent의 SurePrint 원위치 마이크로어레이 합성 플랫폼에서 잉크젯 프린팅 및 플로우 셀 리액터 기술을 사용하였다. 무수 챔버내의 잉크젯 프린팅은 매우 소량의 포스포아라미디트를 2D 평면상의 국한된 결합 영역에 운반할 수 있도록 하여, 수 만개의 염기가 병렬로 첨가되도록 한다. 후속적인 산화 및 디트리틸화(detritylation)는 플로우 셀 리액터에서 수행된다. DNA 합성이 완료되면, 그 다음 올리고뉴클레오타이드는 표면으로부터 쪼개어지고 탈보호된다[25].

[0044] 복수의 DNA 세그먼트 카피가 용이하게 만들어질 수 있도록, 어댑터들은 DNA 세그먼트에 첨가되었다. 어댑터를 갖지 않은 DNA 세그먼트는 DNA 세그먼트의 말단상에 부가적인 기(group)들을 첨가함으로써 다중 카피의 합성을 위한 화학을 "촉진시키기(kick start)"위해 부가적인 화학 공정이 필요할 수 있다.

[0045] 수 천배 과잉의 포스포아미디트 및 활성제 용액을 이용하여 최대 ~99.8% 결합 효율이 달성된다. 마찬가지로, 수백만 배 과잉의 디트리틸화제는 완료에 가까운 5'-히드록실 보호기의 제거를 유도한다. 플로우 셀 리액터에서 조절된 공정은 탈푸린화를 현저히 감소시켰으며, 이는 가장 일반적인 부반응(side reaction)이다[22]. 최대 244000 고유 서열은 병렬로 합성될 수 있으며, ~1-10 피코몰 풀의 올리고로서 운반될 수 있다.

[0046] 동결건조된 올리고의 3가지 시료를 Tris 버퍼에서 4℃에서 밤새 배양하고, 파이펫 및 보텍싱으로 주기적으로 혼합하고, 최종적으로 50℃에서 1시간 동안 5ng/ml의 농도로 배양하였다. 남아있는 불용화 물질로서, 시료들을 매일 2-4회 혼합하면서 4℃에서 5일간 추가로 두었다. 그 다음, 시료들을 50℃에서 1시간 및 68℃에서 10분간 배양하고, Ampure XP 파라마그네틱 비드(Beckman Coulter)에서 잔류 합성 부산물로부터 정제하였으며, 단계 795에서 저장될 수 있었다. 시퀀싱 및 디코딩을 도 8에 나타내었다.

[0047] 복합된 올리고 시료를 단계 810에서 패어드-엔드(paired-end) Illumina PCR 프라이머 및 하이-피델리티 AccuPrime 리전트(Invitrogen), Taq 및 내열성 보조 단백질이 함유된 피로코커스(Pyrococcus) 폴리머라아제의 혼합물을 이용하여 증폭하였다(균일한 A/T 대 G/C 프로세싱을 제공하도록 디자인된 유전자증폭기 조건을 이용항 22 PCR 사이클). 증폭된 산물은 비드 정제되고, Agilent 2100 Bioanalyzer에서 정량화되고, Illumina HiSeq

2000에서 패어드-엔드 방식으로 AYB 소프트웨어를 이용하여 시퀀싱되어 104 염기의 판독을 생성하였다.

[0048] 디지털 정보 디코딩은 다음과 같이 수행되었다. 각 올리고의 중심 91 염기를 양 말단으로부터 단계 820에서 시퀀싱하였으며, 전장(117염기) 올리고의 매우 신속한 산출 및 디자인에 부합되지 않은 시퀀스 판독의 제거는 간단하였다. 시퀀스 판독은 인코딩 공정을 정확히 역전시키는 컴퓨터 소프트웨어를 사용하여 단계 830에서 디코딩되었다. 패리티-검사 트리트가 오류를 나타내거나, 어느 단계에서 분명하게 디코딩되지 못하거나 재구성된 컴퓨터 파일로 할당될 수 없는 시퀀스 판독은 단계 840에서 추가 고려대상으로부터 폐기하였다.

[0049] 디코딩된 모든 파일내에서 대다수의 위치들은 많은 다른 시퀀싱된 DNA 올리고들에서 검출되었으며, DNA 합성 또는 시퀀싱 오류에 의해 야기된 어느 불일치를 해소하기 위해 단계 850에서 단순한 다수 투표법(majority voting)을 사용하였다. 이 절차 860의 완료시, 5개의 본래 컴퓨터 파일(210) 중 4개 파일이 완벽하게 재구성되었다. 5번째 컴퓨터 파일은 어느 시퀀싱된 판독으로부터 복구되지 않은 각각 25염기인 두 부위를 수정하기 위해 수동적인 개입이 요구되었다.

[0050] 단계 850에서 디코딩 중에, (염기-3을 통해 바이트로 디코딩하기 전에) DNA 수준에서 인 실리코에서 재구성된 하나의 파일(궁극적으로 watsoncrick.pdf인 것으로 검출됨)은 시퀀싱된 올리고들 중 어느 하나로부터 복구되지 않은 25염기의 두 부위를 함유하였음을 알게되었다. 인코딩의 오버래핑 세그먼트가 주어지면, 4개의 연속하는 오버래핑 세그먼트 중 어느 하나가 이 위치에 상응하는 염기를 함유할 수 있으므로 각 부위는 4개의 연속하는 세그먼트가 합성 또는 시퀀싱되는데 실패하였음을 나타내었다. 상기 두 부위의 검사는 비-검출된 염기가 하기 20-염기 모티프의 긴 반복부내에 포함되었음을 나타내었다:

[0051] 5' GAGCATCTGCAGATGCTCAT 3'

[0052] 이 모티프의 반복부는 자가-역 상보적 패턴(self-reverse complementary pattern)을 갖는 것을 알게되었다. 이를 도 4에 나타내었다.

[0053] 긴, 자가-역 상보적 DNA 세그먼트는, 그 DNA 세그먼트가 본 명세서에 기재된 방법에 사용된 프로토콜에 사용된 합성 반응에 의한 시퀀싱을 저해할 수 있는 내부 비선형 스템-루프 구조를 형성할 가능성에 기인하여, Illumina 패어드-말단 공정을 이용하여 용이하게 시퀀싱되지 않을 가능성이 있다. 결론적으로, 인 실리코 DNA 시퀀스는 반복 모티프 패턴을 보수하도록 변형된 다음, 후속적인 디코딩 단계를 받았다. 추가적인 문제에 부딪히지 않았으며, 최종 디코딩된 파일은 watsoncrick.pdf 파일과 완벽하게 매칭되었다. 긴 자가-상보적 부위가 어느 디자인된 DNA 세그먼트에도 존재하지 않는 것을 보장하는 코드가 미래에 사용될 수 있다.

[0054] 허프먼 인코딩 스킴의 예(Example of Huffman Coding Scheme)

[0055] 표 1은 바이트값(0-255)을 염기-3로 변환하는데 사용된 예시적인 허프먼 인코딩 스킴의 예를 나타낸 것이다. 고 압축된 정보에 대해, 각 바이트값은 동등하게 빈번하게 나타나야 하며, 바이트당 트리츠의 평균 수는 (239*5 + 17*6)/256 = 5.07이 될 것이다. 바이트당 트리츠의 이론적 최대 수는 log(256)/log(3) = 5.05이다.

표 1

표 1

코드 워드 번호	8-비트 ASCII 캐릭터	바이트 값	베이스 3 코딩 (5 또는 6 트리츠)
0		0	22201
1	U	85	22200
2	™	170	22122
3		127	22121
4	″	253	22120
5	4	52	22112
6	ä	138	22111
7	)	41	22110
8	V	86	22102
9	*	42	22101
10	d	100	22100
11	,	44	22022
12	·	250	22020
13	Ñ	132	22021
14	°	161	22012
15	b	98	22010
16		8	22002
17	"	34	22011
18	[NL]	10	22001
19	ï	149	22000
20	W	87	21222
21		21	21221
22	J	74	21220
23	$	36	21212
24	E	69	21210
25	±	177	21202
26		20	21211
27	’	213	21200

[0056]

28	£	163	21201
29	Â	229	21121
30	ˇ	255	21122
31	≈	197	21120
32	Ő	133	21112
33	˛	252	21110
34		26	21111
35	≠	173	21101
36	ó	151	21102
37	R	82	21100
38	K	75	21022
39	%	37	21021
40	¶	166	21011
41	ø	191	21020
42	X	88	21012
43	?	63	21010
44	D	68	21001
45	ñ	150	21002
46	L	76	21000
47		4	20222
48	ő	154	20221
49	Í	234	20212
50		22	20220
51	¢	162	20211
52	i	105	20210
53	f	102	20202
54	´	171	20201
55	h	104	20200
56	©	169	20122
57	*f*	196	20121
58	–	208	20120

[0057]

59	T	84	20112
60	Ç	130	20111
61	í	146	20102
62	H	72	20110
63		16	20101
64	B	66	20100
65		24	20022
66	j	106	20012
67	fl	223	20020
68	:	58	20021
69	â	137	20011
70	I	73	20010
71	e	101	20001
72	®	168	20002
73	µ	181	12221
74	Ø	175	12222
75	°	251	20000
76	(	40	12220
77	å	140	12212
78		17	12211
79	S	83	12210
80	‹	254	12202
81		240	12201
82	÷	214	12200
83	5	53	12122
84		202	12112
85		25	12121
86		18	12120
87	˜	247	12111
88	Æ	174	12110
89	p	112	12102

[0058]

90	Y	89	12101
91	“	210	12100
92	Ÿ	217	12012
93	¯	248	12020
94	¬	194	12021
95	∂	182	12022
96	P	80	12011
97	O	79	12002
98	√	195	12010
99		12	12001
100	—	209	12000
101	•	165	11222
102	ı	245	11221
103		2	11220
104	Q	81	11212
105	&	38	11211
106	ç	141	11202
107	”	211	11210
108	Ô	239	11200
109	_	95	11201
110	+	43	11122
111	‡	224	11121
112	À	203	11112
113	ë	145	11120
114	ì	147	11110
115		19	11111
116	2	50	11101
117	à	136	11102
118	k	107	11100
119	Ü	134	11022
120	m	109	11021

[0059]

121	ô	153	11020
122	î	148	11002
123	Õ	205	11010
124	‘	212	11011
125	6	54	11012
126	Ò	241	11000
127	ú	156	11001
128	s	115	10222
129	t	116	10221
130	N	78	10220
131	C	67	10211
132	F	70	10212
133	≤	178	10210
134	ü	159	10202
135	é	142	10201
136	\	92	10200
137	0	48	10122
138	Z	90	10120
139	/	218	10121
140	~	126	10112
141	'	39	10111
142	€	219	10102
143	ß	167	10110
144	r	114	10101
145	¨	172	10022
146		14	10100
147	x	120	10020
148	ã	139	10021
149	†	160	10012
150	!	33	10011
151	≥	179	10010

[0060]

152	u	117	10002
153	·	225	10001
154	Å	129	10000
155	∑	183	02222
156	Ê	230	02220
157	#	35	02221
158	]	93	02210
159		6	02211
160		32	02212
161	8	56	02201
162	û	158	02202
163	π	185	02121
164	/	47	02122
165	è	143	02200
166	{	123	02111
167	Ã	204	02120
168	Ú	242	02112
169	o	111	02110
170	g	103	02102
171	l	108	02101
172	[TAB]	9	02100
173	A	65	02022
174	˘	249	02020
175	[CR]	13	02021
176	¥	180	02012
177	,	226	02001
178	ê	144	02002
179		15	02010
180	9	57	02011
181	Ä	128	02000
182	á	135	01220

[0061]

183	Û	243	01221
184	æ	190	01222
185	œ	207	01212
186	M	77	01211
187	-	45	01210
188	[	91	01202
189	¿	192	01201
190	∫	186	01122
191	ÿ	216	01200
192	a	97	01112
193	v	118	01120
194	ˆ	246	01121
195	◊	215	01111
196	3	51	01102
197	Œ	206	01110
198	∏	184	01100
199	„	227	01101
200	È	233	01022
201	Ì	237	01021
202	°	188	01020
203	q	113	01012
204	1	49	01011
205	…	201	01010
206	õ	155	01002
207	fi	222	01000
208	Á	231	01001
209		5	00222
210		27	00221
211	É	131	00212
212	§	164	00220
213		3	00211

[0062]

214	.	46	00210
215	w	119	00201
216		28	00202
217	∞	176	00200
218		23	00122
219	@	64	00121
220	ù	157	00120
221	ª	187	00112
222	Ù	244	00110
223	Ó	238	00111
224	`	96	00102
225	Î	235	00101
226	<	60	00022
227		1	00100
228	n	110	00021
229	»	200	00011
230	›	221	00020
231	c	99	00012
232		31	00010
233	Δ	198	00002
234	¡	193	00001
235	}	125	00000
236	|	124	22222
237	ò	152	22222
238	z	122	22222
239	G	71	222212
240	^	94	222211
241	‹	220	222210
242		29	222202
243	«	199	222201
244	=	61	222200

[0063]

245		11	222122
246	‰	228	222121
247	>	62	222120
248	7	55	222112
249	y	121	222111
250		7	222110
251	-	30	222102
252	Ë	232	222101
253	Ω	189	222100
254	;	59	222021
255	Ï	236	222022

[0064]

[0065] 파일의 인코딩

[0066] 임의의 컴퓨터 파일(210)은 바이트의 스트링 $S_{\emptyset}$로서 표현된다(종종 $\emptyset$과 $2^8 - 1$ 사이의 수, 즉, 세트 {0 ...

255}내의 값으로 해석된다). 스트링 $S_{\emptyset}$는 허프만 코드를 이용하여 인코딩되고, 염기-3으로 변환된다. 이는 캐릭터의 스트링 S_1을 트리트 {Ø, 1, 2}로서 생성한다.

[0067] 스트링 S_1의 길이를 (캐릭터로) 산출하는 함수에 대해 *len*()을 작성하고, $n=len(S_1)$으로 정의한다. 염기-3내에 n을 나타내고, 0을 덧붙여 $len(S_2)=20$이 되도록 트리츠의 스트링 S_2를 생성한다. 스트링 병합 $S_4 = S_1.S_3.S_2$를 형성하며, 여기서 S_3는 $len(S_4)$가 25의 정수 배수가 되도록 많아야 24 제로로 이루어진 스트링이 선택된다.

[0068] S_4는 하기 표에 나타낸 스킴을 이용하여 반복 뉴클레오타이드(nt)가 없는 캐릭터 {A, C, G, T}의 DNA 스트링 S_5로 변환된다. S_4의 처음 트리트는 하기 표의 'A'행을 이용하여 인코딩된다. 각 후속적인 트리트에 대해, 캐릭터들은 앞선 캐릭터 변환에 의해 정의된 행으로부터 취해진다.

표 2

[0069]

기록된 이전 nt	인코딩될 다음 트리트		
	Ø	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

[0070] 표: 반복된 뉴클레오타이드가 없는 것을 보장하는 염기-3의 DNA 인코딩

[0071] 인코딩될 각 트리트 t에 대해, 사용된 앞선 뉴클레오타이드로 표지된 행 및 t 표지된 열을 선택하고, 상응하는 표 셀에서 상기 nt를 사용하여 인코딩한다.

[0072] $N = len(S_5)$을 정의하고, ID를 본래 파일을 식별하고 주어진 실험내에서 고유한 2-트리트 스트링이라 한다(DNA의 혼합을 허용하는 것은 하나의 실험에서 다양한 파일들 $S_{\emptyset}$을 형성한다). S_5를 길이 100 nt의 오버래핑 DNA 세그먼트(240)로 분할하고, 각각의 DNA 세그먼트(240)는 DNA 세그먼트(240)의 앞선 것으로부터 25 nt 만큼 오프셋된다. 이는 ((N/25)-3) DNA 세그먼트(240)가 존재할 것이라는 것을 의미하며, 편의상 인덱싱된 i = Ø ...(N/25)-4 이다. 그 DNA 세그먼트 i는 Fi로 표시되며, S_5의 (DNA) 캐릭터 $25i$ ... 25_{i+99}를 함유한다.

[0073] 각 DNA 세그먼트 F_i는 다음과 같이 더욱 처리된다:

[0074] i가 홀수인 경우, 그 DNA 세그먼트 Fi를 역 상보적(reverse complement)이게 한다.

[0075] $i3$를 i의 염기-3 표현이 되게 하고, $len(i3) = 12$가 되도록 충분한 선행하는 제로(leading zeros)를 덧붙인다. ID 및 $i3$내에서 홀수-위치 트리츠의 합(mod 3), 즉, $ID_1 + i3_1 + i3_3 + i3_5 + i3_7 + i3_9 + i3_{11}$로서, P를 산출한다. (P는 (패리티 비트와 유사한) '패리티 트리트(parity trit)'로서 작용하여 ID 및 i에 대해 인코딩된 정보

에서 오류를 체크한다.)

[0076] 인덱싱 정보(250) 스트링 *IX = ID. i2. P*(2+12+1 = 15 트리츠를 포함함)를 형성한다. F_i의 마지막 캐릭터에 의해 정의된 코드 표 행으로 시작하여, 상기 표에 나타낸 것과 동일한 전략을 사용하여 *IX*의 DNA-인코딩(단계 760) 버전을 F_i에 첨부함으로써 인덱싱된 세그먼트 F'_i를 제공한다.

[0077] A와 T 중에 선택하고, C와 G 중에 선택하여 F'_i에 A 또는 T를 앞에 덧붙이고, C 또는 F를 첨부함으로써 F''_i를 형성하되, 만일 가능하다면 무작위로 하며, 그러나 반복된 뉴클레오타이드가 항상 없도록 한다. 이는 DNA 시퀀싱 중에 역 상보적이게 되지 않은 것으로부터 역 상보적이게 된(단계 240) DNA 세그먼트(240)를 구별하는 것을 보장해 준다. 전자는 G|C로 시작하여 T|A로 끝날 것이며; 후자는 A|T로 시작하여 C|G로 끝날 것이다.

[0078] 세그먼트 F''_1는 단계 790에서 실제 DNA 올리고뉴클레오타이드로 합성되며, 단계 790에서 저장되고, 단계 820에서 시퀀싱을 위해 제공될 수 있다.

[0079] 디코딩

[0080] 디코딩은 단계 720에서의 인코딩의 단순한 반대이며, 길이 117 뉴클레오타이드의 시퀀싱된 DNA 세그먼트(240) F''_1로 시작한다. DNA 시퀀싱 절차 중에(예, PCR 반응 중에) 역 상보화는 프래그먼트가 A|T로 시작하여 C|G로 끝나는지, 아니면 G|C로 시작하여 T|A로 끝나는지 관찰함으로써 후속적인 역전에 대해 식별될 수 있다. 이러한 두 '배향(orientation)' 뉴클레오타이드가 제거되면, 각 DNA 세그먼트(240)의 잔존하는 115 뉴클레오타이드는 처음의 100 '메세지' 뉴클레오타이드 및 나머지 15 '인덱싱 정보(250)' 뉴클레오타이드로 분할될 수 있다. 인덱싱 정보 뉴클레오타이드(250)는 파일 식별자 *ID*와 위치 인덱스 *i3* 및 이에 따라 *i*를 검출하도록 디코딩될 수 있으며, 오류는 패리티 트리트 *P*를 시험함으로써 검출될 수 있다. 위치 인덱싱 정보(250)는 DNA-인코딩된 파일(230)의 재구성을 가능케 하며, 이는 그 다음 상기 인코딩 표의 역전(reverse)을 이용하여 염기-3로 변환된 다음, 주어진 허프만 코드를 이용하여 본래의 바이트로 전환될 수 있다.

[0081] 데이터 저장에 대한 디스커션

[0082] DNA 저장은 통상적인 테이프-기반의 저장 또는 디스크-기반의 저장과는 다른 특성을 갖는다. 본 실시예에서 ~750kB의 정보가 10pmol의 DNA로 합성되어, 약 1 테라바이트/그램의 정보 저장 밀도를 제공하였다. DNA 저장은 전력을 필요로 하지 않으며, 낮잡아 평가하여도 (잠재적으로) 수 천년 동안 실행가능하게 유지된다.

[0083] 또한, DNA 기록 저장은 프라이머쌍에 PCR을 적용한 다음, 그 결과 형성된 DNA 용액을 분액(분할)함으로써 대량 병렬식으로 복사될 수 있다. 시퀀싱 프로세스에서 본 기술의 실현 가능한 입증으로, 본 방법은 수회 수행되었으며, 또한 이는 정보를 대규모로 복사한 다음, 그 정보를 물리적으로 둘 이상의 위치로 전송하기 위해 명백하게 사용될 수 있다. 다중 위치에서 정보의 저장은 어느 기록 저장 스킴에 추가적인 견고함을 제공할 것이며, 시설 간의 매우 큰 규모의 데이터 복사 작업에 대해 그 자체로 유용할 수 있다.

[0084] 본 실시예에서 디코딩 대역폭은 디스크(약 1 테라비트/초) 또는 테이프(140메가비트/초)에 비해 3.4비트/초이었으며, 또한 대기 시간이 높다(본 실시예에서 ~20일). 미래의 시퀀싱 기술은 이러한 팩터들 모두를 향상시킬 것으로 예상된다.

[0085] 본 발명의 DNA-저장 또는 테이프 저장을 이용한 기록 저장의 전체 비용 모델링은, 중요 파라미터가 테이프 저장 기술과 매체 사이의 전이에 대한 빈도 및 고정 비용인 것을 나타낸다. 도 3은 DNA-저장이 비용-효율적인 기간을 보여준다. 상부의 굵은 곡선은 본 발명에서 교시한 바와 같은 DNA 저장이 그 너머로 테이프에 비해 덜 비용적인 손익분기 시간(x-축)을 나타낸다. 이는 테이프 기록 저장이 매 3년마다(f = 1/3) 판독 및 재기록되어야 하며, DNA-저장 합성의 상대적 비용 및 테이프 이송 고정비용(y-축)에 따라 달라지는 것을 가정한 것이다. 하부의 굵은 곡선은 5년에 한 번 테이프 이송에 상응하는 것이다. 상기 하부의 굵은 곡선 아래의 영역은 5년에 한 번 보다 더 자주 이송될 경우에 DNA 저장이 비용 효율적인 경우를 나타내며; 두 굵은 곡선 사이에서, 3 내지 5년에 한 번 이송될 경우에 DNA 저장이 비용 효율적임을 나타내며; 그리고 굵은 곡선 위의 영역은 3년에 한 번 보다 덜 자주 이송될 경우에 테이프가 덜 비싼 것을 나타낸다. 수평의 점선은 125-500(현재 값) 및 12.5-50(DNA 합성 비용이 한 자릿수 감소될 경우에 달성되는)의 테이프 이송에 대한 DNA 합성의 상대적 비용의 범위를 나타낸다. 수직의 점선은 이에 상응하는 손익분기 시간을 나타낸다. 모든 축에 대한 로그 스케일에 주목바란다.

[0086]

[0087] 장기 디지털 기록 보관에 대한 한 문제는 DNA-기반의 스토리지를 더 큰 애플리케이션에 적용하는 방법이다. 정보를 인코딩할 필요가 있는 합성 DNA의 염기의 수는 저장될 정보의 양에 따라 선형으로 증가한다. 또한, 짧은 DNA 세그먼트(240)로부터 완전한 길이의 파일을 재구성하기 위해 필요한 인덱싱 정보를 고려해야한다. 인덱싱 정보(250)는 단지 인덱싱될 DNA 세그먼트 수의 대수(logarithm)와 같이 증가한다. 필요한 합성 DNA의 총 양은 부선형으로(sub-linearly) 증가한다. 각각의 DNA 세그먼트(240)의 큰 부분이 점점 더 인덱싱에 필요하나, 보다 긴 스트링의 합성이 미래에는 가능할 것으로 예측되는 것이 합당하더라도, 데이터 및 인덱싱 정보(250) 모두에 대해 이용가능한 불변 114 뉴클레오타이드의 보존적 통제하에서 상기 스킴의 거동을 모델링하였다.

[0088] 정보의 총 양이 증가함에 따라, 인코딩 효율은 단지 서서히 감소한다(도 5). 상기 실험에서(메가바이트 스케일), 인코딩 스킴은 88% 효율이다. 도 5는 효율이 페타바이트(PB, 10^{15}바이트) 스케일에서 데이터 저장에 대해 >70%을 유지하며, 엑사바이트(EB, 10^{18}바이트) 스케일에서 >65%를 유지하며, 그리고 DNA-기반의 저장은 현재의 글로벌 데이터 용량 보다 10의 몇 승배 더 큰 스케일에서 실현가능하게 유지되는 것을 나타낸다. 또한, 도 5는 데이터 용량이 10의 몇 승배 증가함에 따라 (저장된 단위 정보당) 비용은 단지 서서히 올라가는 것을 보여준다. 최신 기술을 사용하여 이용가능한 합성 DNA 세그먼트(240)의 길이를 고려하면 효율 및 비용 스케일은 한층 더 호의적인 것이다. 저장된 정보의 양이 증가함에 따라, 디코딩은 시퀀싱될 올리고들을 더 많이 필요로 한다. 인코딩된 정보의 바이트당 고정된 디코딩 비용은 각 염기가 보다 적은 횟수로 판독되어 디코딩 오류를 겪게될 가능성이 더 크다는 것을 의미한다. 감소된 시퀀싱 범위가 디코딩된 염기 당 오류율에 미치는 영향을 모델링하기 위한 스케일링 분석의 확장은, 인코딩된 정보의 양이 글로벌 데이터 스케일 이상으로 증가함에 따라 오류율이 단지 매우 서서히 증가하는 것을 밝혔다. 또한, 이는 1,308회의 평균 시퀀싱 범위는 신뢰성있는 디코딩을 위해 필요한 것을 상당히 초과하였음을 나타낸다. 이는 보다 낮은 범위로 실험을 시뮬레이팅하기 위해 $79.6x10^{6}$ 판독-페어로부터 부표본화함으로써 확인되었다.

[0089] 도 5는 10(또는 그 이상)의 인자만큼 범위를 감소시키는 것은 변경되지 않은 디코딩 특성들을 이끌 수 있음을 보여주며, 이는 DNA-저장 방법의 견고성을 또한 나타낸다. DNA-기반의 저장의 적용은 이미 정부 및 역사 기록과 같은 광범위한 접근의 낮은 기대치를 갖는 장기 기록 저장을 위해 경제적으로 실행 가능할 수 있다. 과학적 정황으로 일 예는 CERN의 CASTOR 시스템이며, 이는 강입자충돌기(Large Hadron Collider) 데이터의 총 80 PB를 저장하며, 매년 15 PB로 증가하고 있다. 단지 10%만이 디스크상으로 유지되며, CASTOR는 정기적으로 자기 테이프 포맷들 사이를 이동한다. 보다 오래된 데이터의 기록 저장은 이벤트의 잠재적인 미래의 검증을 필요로 하지만, 접근율은 수집 후 2-3년 후에 상당히 감소한다. 추가 예는 천문학, 의학 및 행성간 탐험에서 발견된다.

[0090] 도 5는 저장된 정보의 양이 증가함에 따라 인코딩 효율 및 비용이 변화하는 것을 보여준다. x-축(대수 스케

일)은 인코딩될 정보의 총 양을 나타낸다. 3 제타바이트(3 ZB, $3x10^{21}$ 바이트) 글로벌 데이터 추정값을 포함하는 공용 데이터 스케일을 나타낸다. 좌측에 대한 y-축 스케일은 데이터 인코딩에 이용가능한 합성 염기들의 비율로서 측정된 인코딩 효율을 나타낸다. 우측에 대한 y-축 스케일은 현재의 합성 비용 수준(실선) 및 두 자릿수 크기 감소의 경우(파선) 양쪽 모두에서 인코딩 비용에 대한 이에 상응하는 영향을 나타낸다.

[0091] 도 6은 샘플링된 오리지널 $79.6x10^{6}$ 판독-페어의 퍼센트로 나타낸 시퀀싱 범위(x-축; 로그 스케일)의 함수로서의 복구된 염기 당 오류율(per-recovered-base error rate)(y-축)을 나타낸다. 하나의 곡선은 인간의 개입 없이 복구된 4개의 파일들을 나타낸다: 오리지널 판독의 ≥2%가 사용된 경우에 오류는 제로이다. 다른 곡선은 본 시험의 이론적 오류율 모델로부터 몬테 카를로(Monte Carlo) 시뮬레이션에 의해 획득된다. 마지막 곡선은 수동적 보정이 필요한 파일(watsoncrick.pdf)을 나타낸다. 최저 가능 오류율은 0.0036%이다. 박스 영역은 표시된 부분을 확대하여 나타낸 것이다.

[0092] 데이터 저장에 부가적으로, 본 발명의 가르침은 또한 스테가노그래피(steganography)에 사용될 수 있다.

[0093] 참고문헌

[0094] 1. Bancroft, C., Bowler, T., Bloom, B. & Clelland, C. T. Long-term storage of information in DNA. *Science* 293, 1763-1765 (2001)

[0095] 2. Cox, J. P. L. Long-term data storage in DNA. *TRENDS Biotech.* 19, 247-250 (2001)

[0096] 3. Baum, E. B. Building an associative memory vastly larger than the brain. *Science* 268, 583-585 (1995)

[0097] 4. Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. *Nature* 399, 533-534 (1999)

[0098] 5. Kac, E. Genesis (1999) http://www.ekac.org/geninfo.html accessed online, 2 April 2012

[0099] 6. Wong, P. C., Wong, K.-K. & Foote, H. Organic data memory. Using the DNA approach. *Comm. ACM* 46, 95-98 (2003)

[0100] 7. Ailenberg, M. & Rotstein, O. D. An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* 47, 747-754 (2009)

[0101] 8. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52-56 (2010)

[0102] 9. MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms.* (Cambridge University Press, 2003)

[0103] 10. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids. *Nature* 171, 737-738 (1953)

[0104] 11. Shapiro, B. *et al.* Rise and fall of the Beringian steppe bison. *Science* 306, 1561-1565 (2004)

[0105] 12. Poinar, H. K. *et al.* Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 311, 392-394 (2005)

[0106] 13. Willerslev, E. *et al.* Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* 317, 111-114 (2007)

[0107] 14. Green, R. E. *et al.* A draft sequence of the Neanderthal genome. *Science* 328, 710-722 (2010)

[0108] 15. Anchordoquy, T. J. & Molina, M. C. Preservation of DNA. *Cell Preservation Tech.* 5, 180-188 (2007)

[0109] 16. Bonnet, J. *et al.* Chain and conformation stability of solid-state DNA: implications for room temperature storage. *Nucl. Acids Res.* 38, 1531-1546 (2010)

[0110] 17. Lee, S. B., Crouse, C. A. & Kline, M. C. Optimizing storage and handling of DNA extracts. *Forensic Sci. Rev.* 22, 131-144 (2010)

[0111] 18. Tsaftaris, S. A. & Katsaggelos, A. K. On designing DNA databases for the storage and retrieval of digital signals. *Lecture Notes Comp. Sci.* 3611, 1192-1201 (2005)

[0112] 19. Yamamoto, M., Kashiwamura, S., Ohuchi, A. & Furukawa, M. Large-scale DNA memory based on the nested PCR. *Natural Computing* 7, 335-346 (2008)

[0113] 20. Kari, L. & Mahalingam, K. DNA computing: a research snapshot. *In* Atallah, M. J. & Blanton, M. (eds.) *Algorithms and Theory of Computation Handbook, vol. 2.* 2nd ed. pp. 31-1-31-24 (Chapman & Hall, 2009)

[0114] 21. Chen, P. M., Lee, E. K., Gibson, G. A., Katz, R. H. & Patterson, D. A. RAID: high-performance, reliable secondary storage. *ACM Computing Surveys* 26, 145-185 (1994)

[0115] 22. Le Proust, E. M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucl. Acids Res.* 38, 2522-2540 (2010)

[0116] 23. Kosuri, S. *et al.* A scalable gene synthesis platform using high-fidelity DNA microchips. *Nature Biotech.* 28, 1295-1299 (2010)

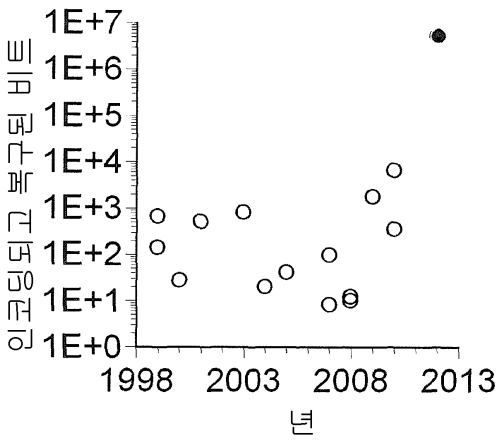
[0117] 24. Beaucage, S. L. & Caruthers, M. H. Deoxynucleoside phosphoramidites - a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* 22, 1859-1862 (1981)

[0118] 25. Cleary, M. A. *et al.* Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nature Methods* 1, 241-248 (2004)

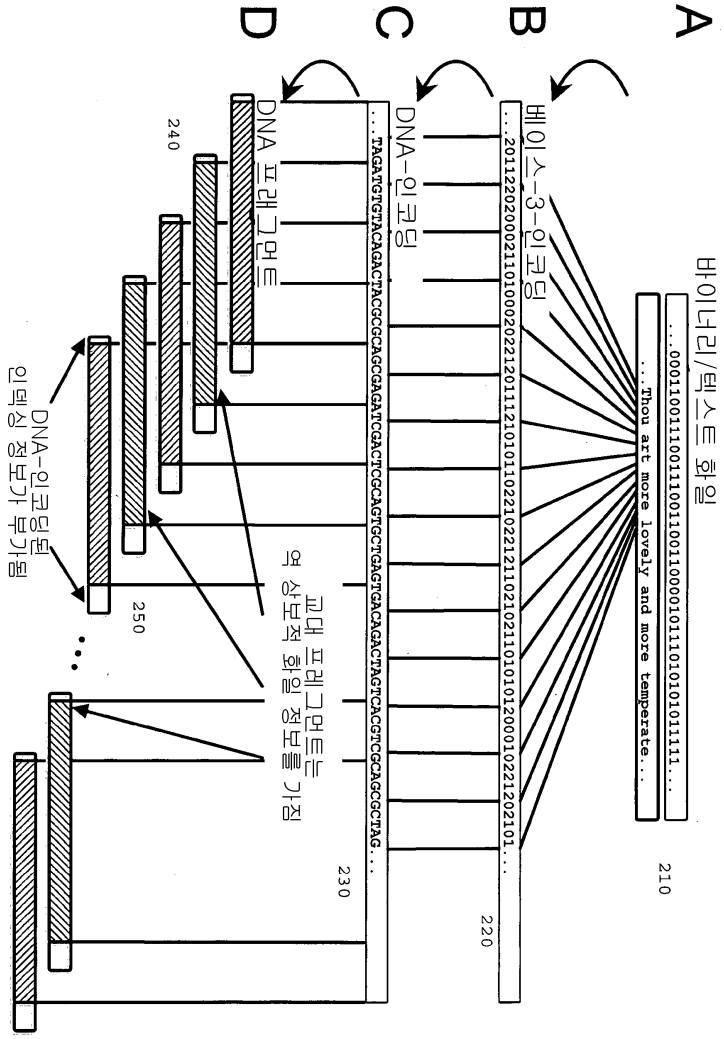
[0119] 26. Aird, D. *et al.* Analysing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, R18 (2011)

도면

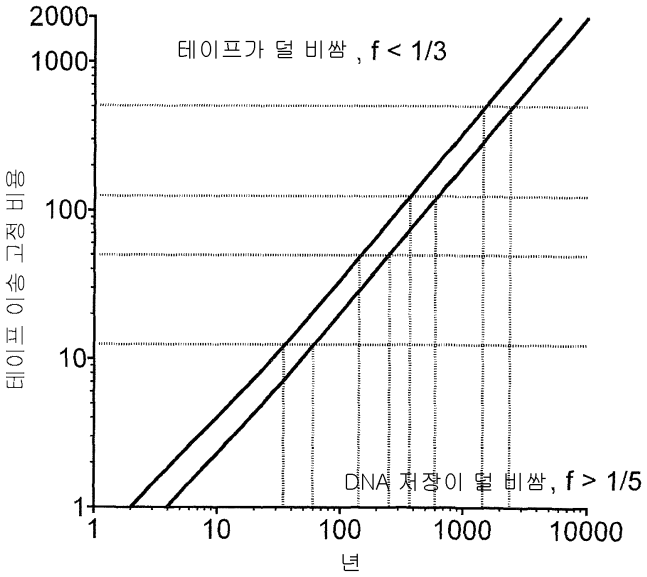
도면1



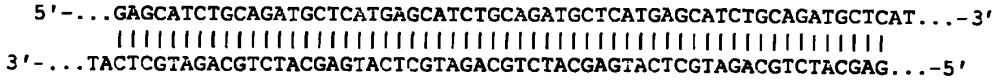
도면2


A
B
C
D
바이너리/텍스트 화일
...0001100111001110011001000010111010101011111...
...Thou art more lovely and more temperate...
210
베이스-3-인 코딩
...2011220200021101000202212011121010110221021211021021101010120001022120210 1...
220
DNA-인 코딩
...TAGATGTGTACAGACTACGCGCAGCGAGATCGACTCGCAGTGCTGAGTGACAGACTAGTCACGTCGCAGCGCTAG...
230
교대 프레그먼트는
옥 상보적 화일 정보를 가짐
DNA 프래그먼트
240
DNA-인코딩된
인덱싱 정보가 부가됨
250

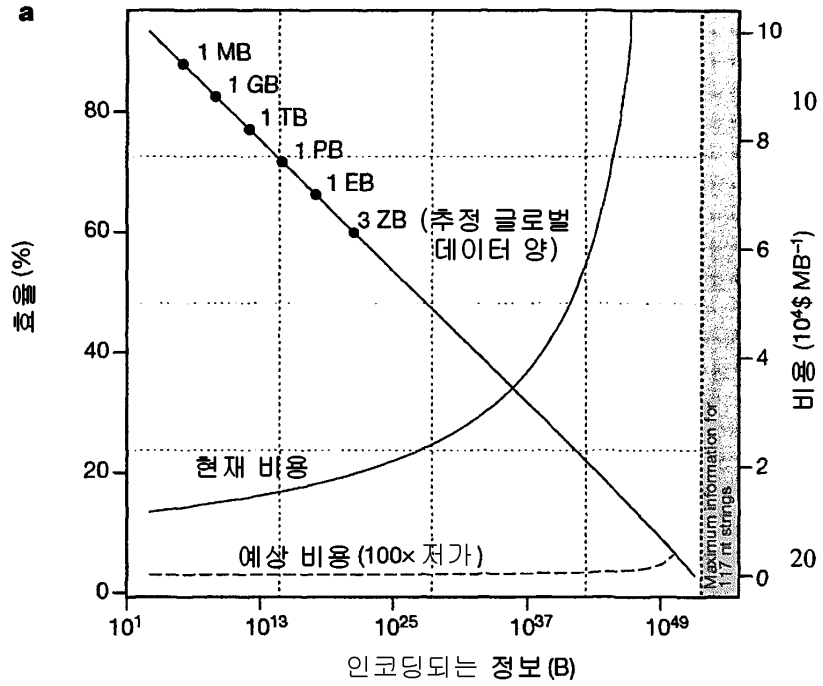
도면3


2000
1000
100
10
1
테이프가 덜 비쌈 , f < 1/3
DNA 저장이 덜 비쌈, f > 1/5
테이프 이송 고정 비용
1
10
100
1000
10000
년

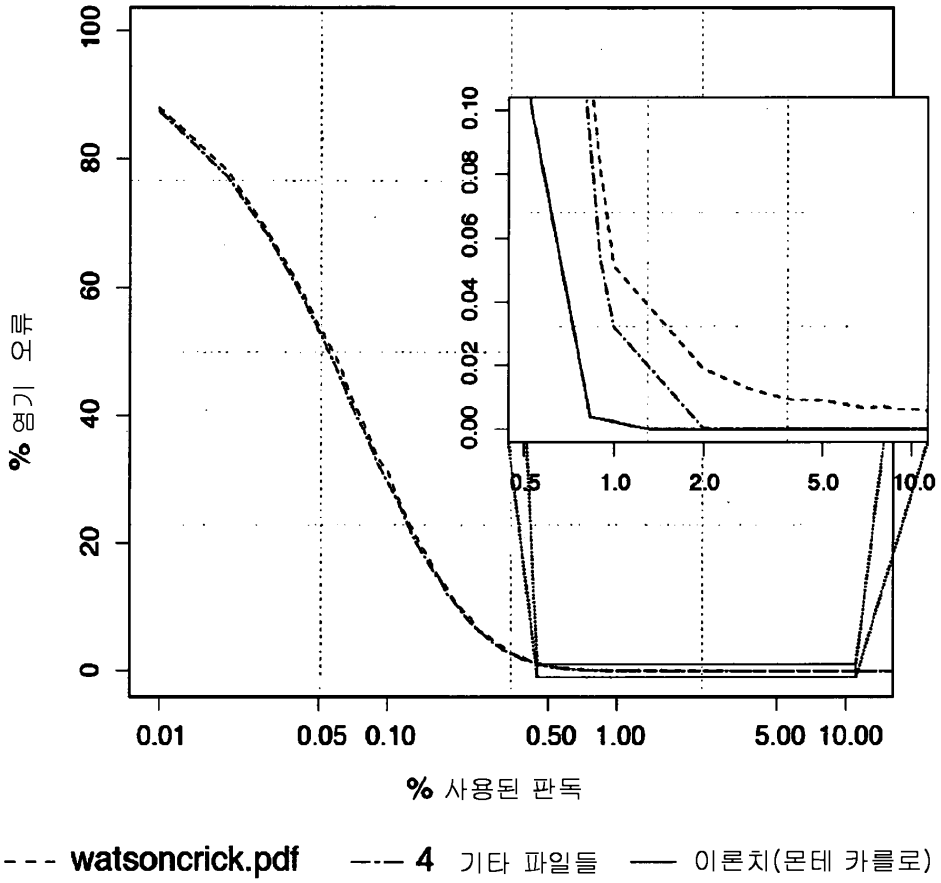
도면4


5'-...GAGCATCTGCAGATGCTCATGAGCATCTGCAGATGCTCATGAGCATCTGCAGATGCTCAT...-3'
3'-...TACTCGTAGACGTCTACGAGTACTCGTAGACGTCTACGAGTACTCGTAGACGTCTACGAG...-5'

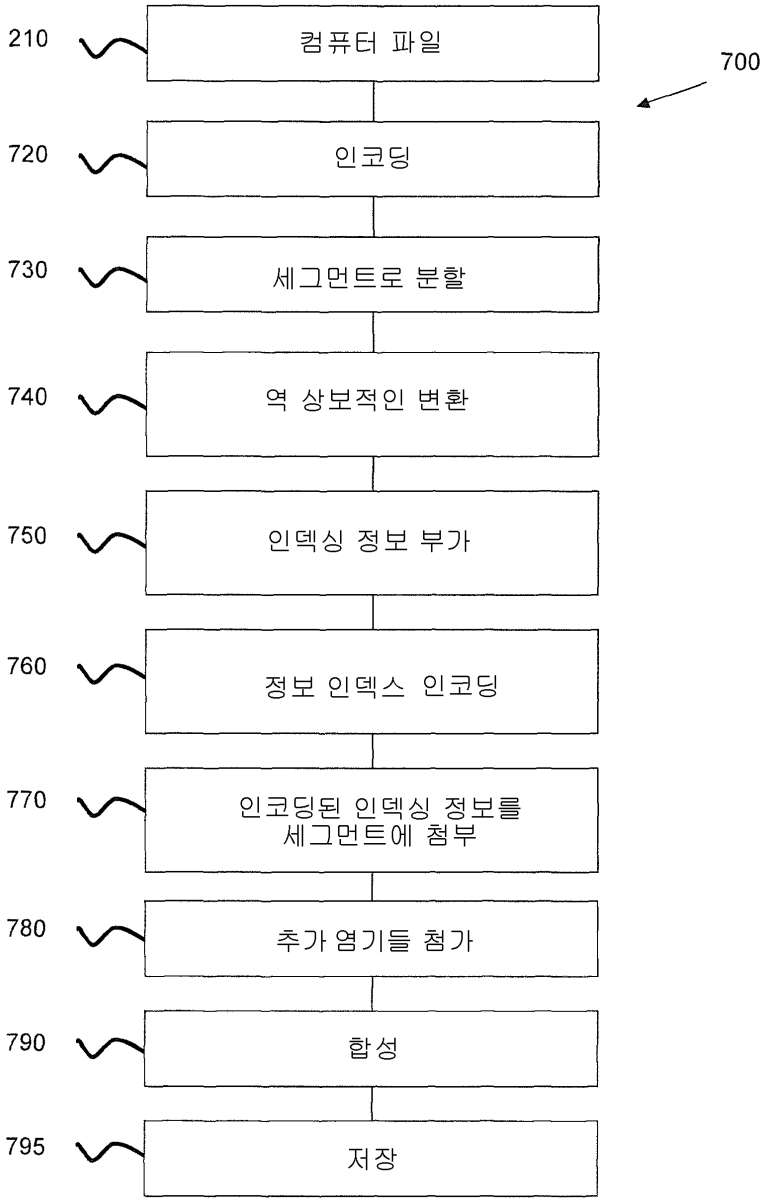
도면5


a
1 MB
1 GB
1 TB
1 PB
1 EB
3 ZB (추정 글로벌 데이터 양)
현재 비용
예상 비용(100× 저가)
Maximum information for 1.17 nt strings
효율 (%)
비용 (10^{-4} MB^{-1})
인코딩되는 정보 (B)
0
20
40
60
80
0
2
4
6
8
10
10
20
10^{1}
10^{13}
10^{25}
10^{37}
10^{49}

도면6


% 영기 오류
100
80
60
40
20
0
0.01
0.05
0.10
0.50
1.00
5.00
10.00
% 사용된 판독
0.00
0.02
0.04
0.06
0.08
0.10
0.5
1.0
2.0
5.0
10.0
--- watsoncrick.pdf
—·— 4 기타 파일들
—— 이론치(몬테 카를로)

도면7


210
컴퓨터 파일
700
720
인코딩
730
세그먼트로 분할
740
역 상보적인 변환
750
인덱싱 정보 부가
760
정보 인덱스 인코딩
770
인코딩된 인덱싱 정보를
세그먼트에 첨부
780
추가 염기들 첨가
790
합성
795
저장

도면8