(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0101760
(43) 공개일자 2023년07월06일

(51) 국제특허분류(Int. Cl.)
G16B 50/50 (2019.01) *G16B 30/00* (2019.01)
H03M 7/30 (2006.01) *H03M 7/40* (2006.01)
(52) CPC특허분류
G16B 50/50 (2019.02)
G16B 30/00 (2019.02)
(21) 출원번호 10-2022-7044606
(22) 출원일자(국제) 2021년11월05일
심사청구일자 없음
(85) 번역문제출일자 2022년12월19일
(86) 국제출원번호 PCT/US2021/058364
(87) 국제공개번호 WO 2022/099097
국제공개일자 2022년05월12일
(30) 우선권주장
63/110,308 2020년11월05일 미국(US)

(71) 출원인
일루미나, 인코포레이티드
미국 캘리포니아 92122 샌디에고 일루미나 웨이 5200
(72) 발명자
리즈크 기욤 알렉산더 파스칼
미국 캘리포니아주 92122 샌디에이고 5200 일루미나 웨이
(74) 대리인
특허법인아주김장리

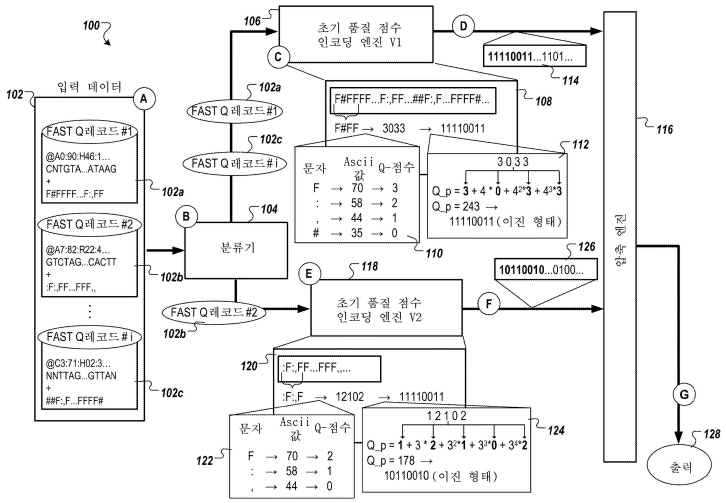
전체 청구항 수 : 총 72 항

(54) 발명의 명칭 **품질 점수 압축**

(57) 요 약

핵산 시퀀스 데이터를 압축하기 위한 방법, 시스템 및 컴퓨터 프로그램. 방법은, (i) 판독 시퀀스(read sequence), 및 (ii) 복수의 품질 점수(quality score)들을 표현하는 핵산 시퀀스 데이터를 획득하는 단계, 판독 시퀀스가 적어도 하나의 N 염기를 포함하는지 여부를 결정하는 단계, 판독 시퀀스가 적어도 하나의 N 염기를 포함하지 않는다는 결정에 기초하여, 기수-(x 마이너스 1) 수(base-(x minus 1) number)를 사용하여 판독 시퀀스의 품질 점수들 각각을 인코딩하기 위해 제1 인코딩 프로세스를 사용함으로써 제1 인코딩된 데이터 세트를 생성하는 단계 - 여기서 x는 핵산 시퀀싱 디바이스에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수임 -, 및 제2 인코딩 프로세스를 사용하여 제1 인코딩된 데이터 세트를 인코딩함으로써, 압축될 데이터를 압축하는 단계를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
H03M 7/3084 (2013.01)
H03M 7/4006 (2013.01)
H03M 7/6011 (2013.01)
H03M 7/6017 (2013.01)
H03M 7/70 (2013.01)

명 세 서

청구범위

청구항 1

핵산 시퀀스 데이터를 압축하기 위한 방법으로서,

하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 단계 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출(base call)들에 대응하는 데이터를 포함하는 판독 시퀀스(read sequence), 및

(ii) 복수의 품질 점수(quality score)들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 생성되었을 가능성을 나타냄 -;

하나 이상의 컴퓨터들에 의해, 상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하는지 여부를 결정하는 단계;

상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하지 않는다는 결정에 기초하여, 하나 이상의 컴퓨터들에 의해, 기수-(x 마이너스 1) 수(base-(x minus 1) number)를 사용하여 상기 판독 시퀀스의 상기 품질 점수들 각각을 인코딩하기 위해 제1 인코딩 프로세스를 사용함으로써 제1 인코딩된 데이터 세트를 생성하는 단계 - x는 상기 핵산 시퀀싱 디바이스에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수임 -; 및

하나 이상의 컴퓨터들에 의해, 제2 인코딩 프로세스를 사용하여 상기 제1 인코딩된 데이터 세트를 인코딩함으로써, 압축될 상기 데이터를 압축하는 단계

를 포함하는, 방법.

청구항 2

제1항에 있어서, x는 3과 같은, 방법.

청구항 3

제2항에 있어서, 상기 제1 인코딩 프로세스는, 하나 이상의 컴퓨터들에 의해, 상기 판독 시퀀스의 상기 복수의 품질 점수들 중 5개의 품질 점수들의 각각의 세트의 각각의 품질 점수를 기수-3 수로서 표현함으로써 상기 5개의 품질 점수들의 세트를 단일 바이트로 인코딩하는 것을 포함하는, 방법.

청구항 4

제1항에 있어서,

상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함한다는 결정에 기초하여, 하나 이상의 컴퓨터들에 의해, 상기 판독 시퀀스의 4개의 품질 점수들의 각각의 세트를 단일 바이트의 메모리로 인코딩하기 위해 제3 인코딩 프로세스를 사용함으로써 제2 인코딩 데이터 세트를 생성하는 단계; 및

하나 이상의 컴퓨터들에 의해, 상기 제2 인코딩 데이터를 인코딩하기 위해 제4 인코딩 프로세스를 사용하는 단계

를 추가로 포함하는, 방법.

청구항 5

제4항에 있어서, 상기 제2 인코딩 프로세스와 상기 제4 인코딩 프로세스는 동일한, 방법.

청구항 6

제1항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 방법.

청구항 7

제1항에 있어서, 상기 제1 인코딩된 데이터 세트는 상기 복수의 품질 점수들의 압축된 버전인, 방법.

청구항 8

제1항에 있어서, 상기 제2 인코딩 프로세스는 압축 프로세스인, 방법.

청구항 9

제8항에 있어서, 상기 압축 프로세스는 범위 인코더(range encoder)의 PPMD(Prediction by Partial Matching) 구현을 포함하는, 방법.

청구항 10

제9항에 있어서, 상기 제1 인코딩된 데이터 세트의 주어진 값에 대해, 상기 주어진 값은 상기 제1 인코딩된 데이터 세트 내의 상기 주어진 값의 위치에 대한 4비트 컨텍스트에 따라 압축되는, 방법.

청구항 11

핵산 시퀀스 데이터를 압축하기 위한 시스템으로서,

하나 이상의 데이터 처리 장치; 및

상기 하나 이상의 데이터 처리 장치에 의해 실행될 때, 상기 하나 이상의 데이터 처리 장치로 하여금 동작들을 수행하게 하는 명령어들을 저장한 하나 이상의 컴퓨터 판독가능 저장 디바이스들

을 포함하되, 상기 동작들은,

하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 것 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및

(ii) 복수의 품질 점수들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 생성되었을 가능성을 나타냄 -;

상기 하나 이상의 컴퓨터들에 의해, 상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하는지 여부를 결정하는 것;

상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하지 않는다는 결정에 기초하여, 상기 하나 이상의 컴퓨터들에 의해, 기수-(x 마이너스 1) 수를 사용하여 상기 판독 시퀀스의 상기 품질 점수들 각각을 인코딩하기 위해 제1 인코딩 프로세스를 사용함으로써 제1 인코딩된 데이터 세트를 생성하는 것 - x는 상기 핵산 시퀀싱 디바이스에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수임 -; 및

상기 하나 이상의 컴퓨터들에 의해, 제2 인코딩 프로세스를 사용하여 상기 제1 인코딩된 데이터 세트를 인코딩함으로써, 압축될 상기 데이터를 압축하는 것

을 포함하는, 시스템.

청구항 12

제11항에 있어서, x는 3과 같은, 시스템.

청구항 13

제12항에 있어서, 상기 제1 인코딩 프로세스는, 상기 하나 이상의 컴퓨터들에 의해, 상기 판독 시퀀스의 상기 복수의 품질 점수들 중 5개의 품질 점수들의 각각의 세트의 각각의 품질 점수를 기수-3 수로서 표현함으로써 상기 5개의 품질 점수들의 세트를 단일 바이트로 인코딩하는 것을 포함하는, 시스템.

청구항 14

제11항에 있어서, 상기 동작들은,

상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함한다는 결정에 기초하여, 상기 하나 이상의 컴퓨터들에 의해, 상기 판독 시퀀스의 4개의 품질 점수들의 각각의 세트를 단일 바이트의 메모리로 인코딩하기 위해 제3 인코딩 프로세스를 사용함으로써 제2 인코딩 데이터 세트를 생성하는 것; 및

상기 하나 이상의 컴퓨터들에 의해, 상기 제2 인코딩 데이터를 인코딩하기 위해 제4 인코딩 프로세스를 사용하는 것

을 추가로 포함하는, 시스템.

청구항 15

제14항에 있어서, 상기 제2 인코딩 프로세스와 상기 제4 인코딩 프로세스는 동일한, 시스템.

청구항 16

제11항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 시스템.

청구항 17

제11항에 있어서, 상기 제1 인코딩된 데이터 세트는 상기 복수의 품질 점수들의 압축된 버전인, 시스템.

청구항 18

제11항에 있어서, 상기 제2 인코딩 프로세스는 압축 프로세스인, 시스템.

청구항 19

제18항에 있어서, 상기 압축 프로세스는 범위 인코더의 PPMD(Prediction by Partial Matching) 구현을 포함하는, 시스템.

청구항 20

제19항에 있어서, 상기 제1 인코딩된 데이터 세트의 주어진 값에 대해, 상기 주어진 값은 상기 제1 인코딩된 데이터 세트 내의 상기 주어진 값의 위치에 대한 4비트 컨텍스트에 따라 압축되는, 시스템.

청구항 21

데이터 처리 장치에 의해 실행될 때, 상기 데이터 처리 장치로 하여금 동작들을 수행하게 하는 명령어들을 저장한 컴퓨터 판독가능 저장 디바이스로서, 상기 동작들은,

핵산 시퀀스 데이터를 획득하는 것 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및

(ii) 복수의 품질 점수들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 생성되었을 가능성을 나타냄 -;

하나 이상의 컴퓨터들에 의해, 상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하는지 여부를 결정하는 것;

상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하지 않는다는 결정에 기초하여, 기수-(x 마이너스 1) 수를 사용하여 상기 판독 시퀀스의 상기 품질 점수들 각각을 인코딩하기 위해 제1 인코딩 프로세스를 사용함으로써 제1 인코딩된 데이터 세트를 생성하는 것 - x는 상기 핵산 시퀀싱 디바이스에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수임 -; 및

제2 인코딩 프로세스를 사용하여 상기 제1 인코딩된 데이터 세트를 인코딩함으로써, 압축될 상기 데이터를 압축하는 것

을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 22

제21항에 있어서, x는 3과 같은, 컴퓨터 판독가능 저장 디바이스.

청구항 23

제22항에 있어서, 상기 제1 인코딩 프로세스는, 상기 판독 시퀀스의 상기 복수의 품질 점수들 중 5개의 품질 점수들의 각각의 세트의 각각의 품질 점수를 기수-3 수로서 표현함으로써 상기 5개의 품질 점수들의 세트를 단일 바이트로 인코딩하는 것을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 24

제21항에 있어서, 상기 동작들은,

상기 판독 시퀀스가 적어도 하나의 "N" 염기를 포함한다는 결정에 기초하여, 상기 판독 시퀀스의 4개의 품질 점수들의 각각의 세트를 단일 바이트의 메모리로 인코딩하기 위해 제3 인코딩 프로세스를 사용함으로써 제2 인코딩 데이터 세트를 생성하는 것; 및

상기 제2 인코딩 데이터를 인코딩하기 위해 제4 인코딩 프로세스를 사용하는 것

을 추가로 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 25

제24항에 있어서, 상기 제2 인코딩 프로세스와 상기 제4 인코딩 프로세스는 동일한, 컴퓨터 판독가능 저장 디바이스.

청구항 26

제21항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 27

제21항에 있어서, 상기 제1 인코딩된 데이터 세트는 상기 복수의 품질 점수들의 압축된 버전인, 컴퓨터 판독가능 저장 디바이스.

청구항 28

제21항에 있어서, 상기 제2 인코딩 프로세스는 압축 프로세스인, 컴퓨터 판독가능 저장 디바이스.

청구항 29

제28항에 있어서, 상기 압축 프로세스는 범위 인코더의 PPMD(Prediction by Partial Matching) 구현을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 30

제29항에 있어서, 상기 제1 인코딩된 데이터 세트의 주어진 값에 대해, 상기 주어진 값은 상기 제1 인코딩된 데이터 세트 내의 상기 주어진 값의 위치에 대한 4비트 컨텍스트에 따라 압축되는, 컴퓨터 판독가능 저장 디바이스.

청구항 31

핵산 시퀀스 데이터를 압축하기 위한 방법으로서,

하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 단계 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및

(ii) 복수의 품질 점수들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 생성되었을 가능성을 나타냄 -;

하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들 내의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 단계 - 각각의 품질 점수 그룹은 상기 복수의 품질 점수들의 품질 점수들의 서브세트를 포함함 -;

상기 복수의 품질 점수들의 제1 서브세트 내의 각각의 특정 품질 점수에 대해:

하나 이상의 컴퓨터들에 의해, 상기 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라고 결정하는 단계;

상기 품질 점수가 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라는 결정에 기초하여, 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 그룹 매핑을 사용하여, 축소된 시퀀스(reduced sequence) 내의 단일 엔트리로서 사용될 제1 데이터를 생성하는 단계 - 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제1 데이터는 상기 특정 품질 점수 그룹을 표현함 -; 및

하나 이상의 컴퓨터들에 의해, 상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제1 데이터를 집계함으로써 상기 축소된 시퀀스를 생성하는 단계

를 포함하는, 방법.

청구항 32

제31항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 방법.

청구항 33

제31항에 있어서, 상기 복수의 품질 점수들 내의 각각의 품질 점수는 상기 품질 점수에 대한 ASCII 값을 표현하는 데이터인, 방법.

청구항 34

제31항에 있어서, 상기 방법은,

상기 복수의 품질 점수들의 제2 서브세트 내의 각각의 특정 품질 점수에 대해:

하나 이상의 컴퓨터들에 의해, 상기 품질 점수들의 시퀀스의 상기 제2 서브세트 내의 상기 특정 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아니라고 결정하는 단계; 및

하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 단일 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터를 생성하는 단계

를 추가로 포함하되, 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제2 데이터는 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수를 표현하고, 상기 미리 결정된 단일 매핑은 복수의 단일 품질 점수들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 방법.

청구항 35

제34항에 있어서,

하나 이상의 컴퓨터들에 의해, 상기 축소된 시퀀스를 생성하는 단계는,

하나 이상의 컴퓨터들에 의해, 상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제1 데이터를 집계하는 단계; 및

하나 이상의 컴퓨터들에 의해, 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제2 데이터를 집계하는 단계

를 포함하는, 방법.

청구항 36

제31항에 있어서, 하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들에서 복수의 품질 점수 그룹들을 식별하는 단계를 추가로 포함하는, 방법.

청구항 37

제31항에 있어서, 상기 미리 결정된 그룹 매핑은 복수의 상이한 품질 점수 그룹들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 방법.

청구항 38

핵산 시퀀스 데이터를 압축하기 위한 시스템으로서,

하나 이상의 데이터 처리 장치; 및

상기 하나 이상의 데이터 처리 장치에 의해 실행될 때, 상기 하나 이상의 데이터 처리 장치로 하여금 동작들을 수행하게 하는 명령어들을 저장한 하나 이상의 컴퓨터 판독가능 저장 디바이스들

을 포함하되, 상기 동작들은,

하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 것 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및

(ii) 복수의 품질 점수들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 생성되었을 가능성을 나타냄 -;

상기 하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 것 - 각각의 품질 점수 그룹은 상기 복수의 품질 점수들의 품질 점수들의 서브세트를 포함함 -;

상기 복수의 품질 점수들의 제1 서브세트 내의 각각의 특정 품질 점수에 대해:

상기 하나 이상의 컴퓨터들에 의해, 상기 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라고 결정하는 것;

상기 품질 점수가 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라는 결정에 기초하여, 상기 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 그룹 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제1 데이터를 생성하는 것 - 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제1 데이터는 상기 특정 품질 점수 그룹을 표현함 -; 및

상기 하나 이상의 컴퓨터들에 의해, 상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제1 데이터를 집계함으로써 상기 축소된 시퀀스를 생성하는 것

을 포함하는, 시스템.

청구항 39

제38항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 시스템.

청구항 40

제38항에 있어서, 상기 복수의 품질 점수들 내의 각각의 품질 점수는 상기 품질 점수에 대한 ASCII 값을 표현하는 데이터인, 시스템.

청구항 41

제38항에 있어서, 상기 동작들은,

상기 복수의 품질 점수들의 제2 서브세트 내의 각각의 특정 품질 점수에 대해:

상기 하나 이상의 컴퓨터들에 의해, 상기 품질 점수들의 시퀀스의 상기 제2 서브세트 내의 상기 특정 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아니라고 결정하는 것; 및

상기 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 단일 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터를 생성하는 것

을 추가로 포함하되, 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제2 데이터는 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수를 표현하고, 상기 미리 결정된 단일 매핑은 복수의 단일 품질 점수들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 시스템.

청구항 42

제41항에 있어서,

상기 하나 이상의 컴퓨터들에 의해, 상기 축소된 시퀀스를 생성하는 것은,

상기 하나 이상의 컴퓨터들에 의해, 상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제1 데이터를 집계하는 것; 및

상기 하나 이상의 컴퓨터들에 의해, 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제2 데이터를 집계하는 것

을 포함하는, 시스템.

청구항 43

제38항에 있어서, 상기 하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들에서 복수의 품질 점수 그룹들을 식별하는 것을 추가로 포함하는, 시스템.

청구항 44

제38항에 있어서, 상기 미리 결정된 그룹 매핑은 복수의 상이한 품질 점수 그룹들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 시스템.

청구항 45

데이터 처리 장치에 의해 실행될 때, 상기 데이터 처리 장치로 하여금 동작들을 수행하게 하는 명령어들을 저장한 컴퓨터 판독가능 저장 디바이스로서, 상기 동작들은,

핵산 시퀀스 데이터를 획득하는 것 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및

(ii) 복수의 품질 점수들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 생성되었을 가능성을 나타냄 -;

상기 복수의 품질 점수들의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 것 - 각각의 품질 점수 그룹은 상기 복수의 품질 점수들의 품질 점수들의 서브세트를 포함함 -;

상기 복수의 품질 점수들의 제1 서브세트 내의 각각의 특정 품질 점수에 대해:

상기 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라고 결정하는 것;

상기 품질 점수가 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라는 결정에 기초하여, 미리 결정된 그룹 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제1 데이터를 생성하는 것 - 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제1 데이터는 상기 특정 품질 점수 그룹을 표현함 -; 및

상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제1 데이터를 집계함으로써 상기 축소된 시퀀스를 생성하는 것

을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 46

제45항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 47

제45항에 있어서, 상기 복수의 품질 점수들 내의 각각의 품질 점수는 상기 품질 점수에 대한 ASCII 값을 표현하는 데이터인, 컴퓨터 판독가능 저장 디바이스.

청구항 48

제45항에 있어서, 상기 동작들은,

상기 복수의 품질 점수들의 제2 서브세트 내의 각각의 특정 품질 점수에 대해:

상기 품질 점수들의 시퀀스의 상기 제2 서브세트 내의 상기 특정 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아니라고 결정하는 것; 및

미리 결정된 단일 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터를 생성하는 것

을 추가로 포함하되, 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제2 데이터는 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수를 표현하고, 상기 미리 결정된 단일 매핑은 복수의 단일 품질 점수들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 컴퓨터 판독가능 저장 디바이스.

청구항 49

제48항에 있어서,

상기 축소된 시퀀스를 생성하는 것은,

상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제1 데이터를 집계하는 것; 및

상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제2 데이터를 집계하는 것

을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 50

제45항에 있어서, 상기 동작들은 상기 복수의 품질 점수들에서 복수의 품질 점수 그룹들을 식별하는 것을 추가로 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 51

제45항에 있어서, 상기 미리 결정된 그룹 매핑은 복수의 상이한 품질 점수 그룹들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 컴퓨터 판독가능 저장 디바이스.

청구항 52

핵산 시퀀스 데이터를 압축하기 위한 방법으로서,

하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 단계 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및

(ii) 복수의 품질 점수들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 식별되었을 가능성을 나타냄 -;

하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들 내의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 단계 - 각각의 품질 점수 그룹은 상기 복수의 품질 점수들의 품질 점수들의 서브세트를 포함함 -;

상기 복수의 품질 점수들의 제1 서브세트 내의 각각의 특정 품질 점수에 대해:

하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들의 상기 제1 서브세트 내의 상기 특정 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아니라고 결정하는 단계; 및

하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 단일 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제1 데이터를 생성하는 단계 - 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제1 데이터는 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수를 표현함 -; 및

하나 이상의 컴퓨터들에 의해, 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제1 데이터를 집계함으로써 상기 축소된 시퀀스를 생성하는

단계

를 포함하는, 방법.

청구항 53

제52항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 방법.

청구항 54

제52항에 있어서, 상기 복수의 품질 점수들 내의 각각의 품질 점수는 상기 품질 점수에 대한 ASCII 값을 표현하는 데이터인, 방법.

청구항 55

제52항에 있어서, 상기 방법은,

상기 복수의 품질 점수들의 제2 서브세트 내의 각각의 특정 품질 점수에 대해:

하나 이상의 컴퓨터들에 의해, 상기 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라고 결정하는 단계; 및

상기 품질 점수가 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라는 결정에 기초하여, 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 그룹 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터를 생성하는 단계

를 추가로 포함하되, 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제2 데이터는 상기 특정 품질 점수 그룹을 표현하고, 상기 미리 결정된 그룹 매핑은 복수의 상이한 품질 점수 그룹들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 방법.

청구항 56

제55항에 있어서,

하나 이상의 컴퓨터들에 의해, 상기 축소된 시퀀스를 생성하는 단계는,

하나 이상의 컴퓨터들에 의해, 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제1 데이터를 집계하는 단계; 및

하나 이상의 컴퓨터들에 의해, 상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제2 데이터를 집계하는 단계

를 포함하는, 방법.

청구항 57

제52항에 있어서,

하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들에서 복수의 품질 점수 그룹들을 식별하는 단계를 추가로 포함하는, 방법.

청구항 58

제52항에 있어서,

상기 미리 결정된 단일 매핑은 복수의 단일 품질 점수들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 방법.

청구항 59

핵산 시퀀스 데이터를 압축하기 위한 시스템으로서,

하나 이상의 데이터 처리 장치; 및

상기 하나 이상의 데이터 처리 장치에 의해 실행될 때, 상기 하나 이상의 데이터 처리 장치로 하여금 동작들을 수행하게 하는 명령어들을 저장한 하나 이상의 컴퓨터 판독가능 저장 디바이스들

을 포함하되, 상기 동작들은,

하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 것 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및

(ii) 복수의 품질 점수들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 식별되었을 가능성을 나타냄 -;

상기 하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 것 - 각각의 품질 점수 그룹은 상기 복수의 품질 점수들의 품질 점수들의 서브세트를 포함함 -;

상기 복수의 품질 점수들의 제1 서브세트 내의 각각의 특정 품질 점수에 대해:

상기 하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들의 상기 제1 서브세트 내의 상기 특정 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아니라고 결정하는 것; 및

상기 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 단일 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제1 데이터를 생성하는 것 - 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제1 데이터는 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수를 표현함 -; 및

상기 하나 이상의 컴퓨터들에 의해, 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제1 데이터를 집계함으로써 상기 축소된 시퀀스를 생성하는 것

을 포함하는, 시스템.

청구항 60

제59항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 시스템.

청구항 61

제59항에 있어서, 상기 복수의 품질 점수들 내의 각각의 품질 점수는 상기 품질 점수에 대한 ASCII 값을 표현하는 데이터인, 시스템.

청구항 62

제59항에 있어서, 상기 동작들은,

상기 복수의 품질 점수들의 제2 서브세트 내의 각각의 특정 품질 점수에 대해:

상기 하나 이상의 컴퓨터들에 의해, 상기 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라고 결정하는 것; 및

상기 품질 점수가 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라는 결정에 기초하여, 상기 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 그룹 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터를 생성하는 것

을 추가로 포함하되, 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제2 데이터는 상기 특정 품질 점수 그룹을 표현하고, 상기 미리 결정된 그룹 매핑은 복수의 상이한 품질 점수 그룹들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 시스템.

청구항 63

제62항에 있어서,

상기 하나 이상의 컴퓨터들에 의해, 상기 축소된 시퀀스를 생성하는 것은,

상기 하나 이상의 컴퓨터들에 의해, 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제1 데이터를 집계하는 것; 및

상기 하나 이상의 컴퓨터들에 의해, 상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제2 데이터를 집계하는 것

을 포함하는, 시스템.

청구항 64

제59항에 있어서, 상기 시스템은,

상기 하나 이상의 컴퓨터들에 의해, 상기 복수의 품질 점수들에서 복수의 품질 점수 그룹들을 식별하는 것을 추가로 포함하는, 시스템.

청구항 65

제59항에 있어서,

상기 미리 결정된 단일 매핑은 복수의 단일 품질 점수들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 시스템.

청구항 66

데이터 처리 장치에 의해 실행될 때, 상기 데이터 처리 장치로 하여금 동작들을 수행하게 하는 명령어들을 저장한 컴퓨터 판독가능 저장 디바이스로서, 상기 동작들은,

핵산 시퀀스 데이터를 획득하는 것 - 상기 핵산 시퀀스 데이터는,

(i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및

(ii) 복수의 품질 점수들을 표현하며, 상기 복수의 품질 점수들의 각각의 품질 점수는 상기 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 식별되었을 가능성을 나타냄 -;

상기 복수의 품질 점수들의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 것 - 각각의 품질 점수 그룹은 상기 복수의 품질 점수들의 품질 점수들의 서브세트를 포함함 -;

상기 복수의 품질 점수들의 제1 서브세트 내의 각각의 특정 품질 점수에 대해:

상기 복수의 품질 점수들의 상기 제1 서브세트 내의 상기 특정 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아니라고 결정하는 것; 및

하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 단일 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제1 데이터를 생성하는 것 - 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제1 데이터는 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수를 표현함 -; 및

하나 이상의 컴퓨터들에 의해, 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제1 데이터를 집계함으로써 상기 축소된 시퀀스를 생성하는 것

을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 67

제66항에 있어서, 상기 획득된 데이터는 FASTQ 파일을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 68

제66항에 있어서, 상기 복수의 품질 점수들 내의 각각의 품질 점수는 상기 품질 점수에 대한 ASCII 값을 표현하는 데이터인, 컴퓨터 판독가능 저장 디바이스.

청구항 69

제66항에 있어서, 상기 동작들은,

상기 복수의 품질 점수들의 제2 서브세트 내의 각각의 특정 품질 점수에 대해:

상기 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라고 결정하는 것; 및

상기 품질 점수가 상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라는 결정에 기초하여, 미리 결정된 그룹 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터를 생성하는 것

을 추가로 포함하되, 상기 축소된 시퀀스 내의 상기 단일 엔트리로서 사용될 상기 제2 데이터는 상기 특정 품질 점수 그룹을 표현하고, 상기 미리 결정된 그룹 매핑은 복수의 상이한 품질 점수 그룹들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 컴퓨터 판독가능 저장 디바이스.

청구항 70

제69항에 있어서,

상기 축소된 시퀀스를 생성하는 것은,

상기 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 상기 생성된 제1 데이터를 집계하는 것; 및

상기 특정 품질 점수 그룹들 각각에 대해 상기 생성된 제2 데이터를 집계하는 것

을 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 71

제66항에 있어서, 상기 동작들은,

상기 복수의 품질 점수들에서 복수의 품질 점수 그룹들을 식별하는 것을 추가로 포함하는, 컴퓨터 판독가능 저장 디바이스.

청구항 72

제66항에 있어서,

상기 미리 결정된 단일 매핑은 복수의 단일 품질 점수들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의하는, 컴퓨터 판독가능 저장 디바이스.

발명의 설명

기 술 분 야

[0001] 관련 출원의 상호 참조

[0002] 본 출원은 2020년 11월 5일자로 출원된 미국 출원 제63/110,308호의 이익을 주장하며, 이의 전체 내용은 본 명세서에 참고로 포함된다.

배 경 기 술

[0003] 일부 경우들에서, 게놈 시퀀싱은 게놈 데이터의 뉴클레오티드 또는 다른 구성요소 부분들을 식별하는 방법을 설명한다. 컴퓨터는 게놈 데이터의 하나 이상의 세트들을 분석하고 뉴클레오티드와 같은 구성요소 부분들의 집합을 주어진 참조 게놈 내의 각자의 위치들과 상관시키는 데 사용될 수 있다. 이러한 방식으로, 컴퓨터는 분자 마커들의 집합을 참조 게놈 상으로 "매핑"할 수 있다.

발명의 내용

[0004] 일반적으로, 본 개시내용은 게놈 데이터에 기초하여 시퀀싱 엔진에 의해 생성된 품질 점수(quality score)들의 압축을 위한 방법, 시스템 및 컴퓨터 프로그램에 관한 것이다. 일 구현예에서, 게놈 데이터에 기초하여 시퀀싱

엔진에 의해 생성된 품질 점수들은 초기에, 하나 이상의 품질 점수들을, 품질 점수들의 세트를 표현하는 데이터 항목들의 시퀀스 내의 단일 데이터 항목으로 그룹화함으로써 압축될 수 있다. 이어서, 데이터 항목들의 시퀀스는 최종 압축된 형태로 추가로 압축되거나 인코딩될 수 있다.

[0005] 본 개시내용의 하나의 혁신적인 양태에 따르면, 핵산 시퀀스 데이터를 압축하기 위한 방법이 개시된다. 일 양태에서, 본 방법은, 하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 단계 - 핵산 시퀀스 데이터는 (i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출(base call)들에 대응하는 데이터를 포함하는 판독 시퀀스(read sequence), 및 (ii) 복수의 품질 점수들을 표현하며, 복수의 품질 점수들의 각각의 품질 점수는 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 생성되었을 가능성을 나타냄 -, 하나 이상의 컴퓨터들에 의해, 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하는지 여부를 결정하는 단계, 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하지 않는다는 결정에 기초하여, 하나 이상의 컴퓨터들에 의해, 기수-(x 마이너스 1) 수(base-(x minus 1) number)를 사용하여 판독 시퀀스의 품질 점수들 각각을 인코딩하기 위해 제1 인코딩 프로세스를 사용함으로써 제1 인코딩된 데이터 세트를 생성하는 단계 - 여기서 x는 핵산 시퀀싱 디바이스에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수임 -, 및 하나 이상의 컴퓨터들에 의해, 제2 인코딩 프로세스를 사용하여 제1 인코딩된 데이터 세트를 인코딩함으로써, 압축될 데이터를 압축하는 단계를 포함할 수 있다.

[0006] 다른 버전들은 컴퓨터 판독가능 저장 디바이스들 상에 인코딩된 명령어들에 의해 정의되는 방법들의 액션들을 수행하기 위한 대응하는 시스템들, 장치, 및 컴퓨터 프로그램들을 포함한다.

[0007] 이들 및 다른 버전들은 선택적으로 하기의 특징들 중 하나 이상을 포함할 수 있다. 예를 들어, 일부 구현예들에서, x는 3과 같다.

[0008] 일부 구현예들에서, 제1 인코딩 프로세스는, 하나 이상의 컴퓨터들에 의해, 판독 시퀀스의 복수의 품질 점수들 중 5개의 품질 점수들의 각각의 세트의 각각의 품질 점수를 기수-3 수로서 표현함으로써 5개의 품질 점수들의 세트를 단일 바이트로 인코딩하는 것을 포함할 수 있다.

[0009] 일부 구현예들에서, 본 방법은, 판독 시퀀스가 적어도 하나의 "N" 염기를 포함한다는 결정에 기초하여, 하나 이상의 컴퓨터들에 의해, 판독 시퀀스의 4개의 품질 점수들의 각각의 세트를 단일 바이트의 메모리로 인코딩하기 위해 제3 인코딩 프로세스를 사용함으로써 제2 인코딩 데이터 세트를 생성하는 단계, 및 하나 이상의 컴퓨터들에 의해, 제2 인코딩 데이터를 인코딩하기 위해 제4 인코딩 프로세스를 사용하는 단계를 추가로 포함할 수 있다.

[0010] 일부 구현예들에서, 제2 인코딩 프로세스와 제4 인코딩 프로세스는 동일하다.

[0011] 일부 구현예들에서, 획득된 데이터는 FASTQ 파일을 포함할 수 있다.

[0012] 일부 구현예들에서, 제1 인코딩된 데이터 세트는 복수의 품질 점수들의 압축된 버전이다.

[0013] 일부 구현예들에서, 제2 인코딩 프로세스는 압축 프로세스이다.

[0014] 일부 구현예들에서, 압축 프로세스는 범위 인코더(range encoder)의 PPMD(Prediction by Partial Matching) 구현을 포함한다.

[0015] 일부 구현예들에서, 제1 인코딩된 데이터 세트의 주어진 값에 대해, 주어진 값은 제1 인코딩된 데이터 세트 내의 주어진 값의 위치에 대한 4비트 컨텍스트에 따라 압축된다.

[0016] 본 개시내용의 다른 혁신적인 양태에 따르면, 핵산 시퀀스 데이터를 압축하기 위한 다른 방법이 개시된다. 일 양태에서, 본 방법은, 하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 단계 - 핵산 시퀀스 데이터는 (i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및 (ii) 복수의 품질 점수들을 표현하며, 복수의 품질 점수들의 각각의 품질 점수는 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 생성되었을 가능성을 나타냄 -, 하나 이상의 컴퓨터들에 의해, 복수의 품질 점수들의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 단계 - 각각의 품질 점수 그룹은 복수의 품질 점수들의 품질 점수들의 서브세트를 포함함 -, 복수의 품질 점수들의 제1 서브세트 내의 각각의 특정 품질 점수에 대해: 하나 이상의 컴퓨터들에 의해, 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라고 결정하는 단계, 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라는 결정에 기초하여, 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 그룹 매핑을 사용하여, 축소된 시퀀스(reduced sequence) 내의 단일 엔트리로서 사용될 제1 데이터를 생성하는 단계

- 축소된 시퀀스 내의 단일 엔트리로서 사용될 제1 데이터는 특정 품질 점수 그룹을 표현함 -, 및 하나 이상의 컴퓨터들에 의해, 특정 품질 점수 그룹들 각각에 대해 생성된 제1 데이터를 집계함으로써 축소된 시퀀스를 생성하는 단계를 포함할 수 있다.

[0017] 다른 버전들은 컴퓨터 판독가능 저장 디바이스들 상에 인코딩된 명령어들에 의해 정의되는 방법들의 액션들을 수행하기 위한 대응하는 시스템들, 장치, 및 컴퓨터 프로그램들을 포함한다.

[0018] 이들 및 다른 버전들은 선택적으로 하기의 특징들 중 하나 이상을 포함할 수 있다. 예를 들어, 일부 구현예들에서, 획득된 데이터는 FASTQ 파일을 포함한다.

[0019] 일부 구현예들에서, 복수의 품질 점수들 내의 각각의 품질 점수는 품질 점수에 대한 ASCII 값을 표현하는 데이터이다.

[0020] 일부 구현예들에서, 본 방법은, 복수의 품질 점수들의 제2 서브세트 내의 각각의 특정 품질 점수에 대해: 하나 이상의 컴퓨터들에 의해, 품질 점수들의 시퀀스의 제2 서브세트 내의 특정 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아니라고 결정하는 단계, 및 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 단일 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터를 생성하는 단계를 추가로 포함할 수 있으며, 여기서 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터는 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수를 표현하고, 미리 결정된 단일 매핑은 복수의 단일 품질 점수들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의한다.

[0021] 일부 구현예들에서, 하나 이상의 컴퓨터들에 의해, 축소된 시퀀스를 생성하는 단계는, 하나 이상의 컴퓨터들에 의해, 특정 품질 점수 그룹들 각각에 대해 생성된 제1 데이터를 집계하는 단계, 및 하나 이상의 컴퓨터들에 의해, 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 생성된 제2 데이터를 집계하는 단계를 포함할 수 있다.

[0022] 일부 구현예들에서, 본 방법은, 하나 이상의 컴퓨터들에 의해, 복수의 품질 점수들에서 복수의 품질 점수 그룹들을 식별하는 단계를 추가로 포함할 수 있다.

[0023] 일부 구현예들에서, 미리 결정된 그룹 매핑은 복수의 상이한 품질 점수 그룹들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의한다.

[0024] 본 개시내용의 다른 혁신적인 양태에 따르면, 핵산 시퀀스 데이터를 압축하기 위한 다른 방법이 개시된다. 일 양태에서, 본 방법은, 하나 이상의 컴퓨터들에 의해, 핵산 시퀀스 데이터를 획득하는 단계 - 핵산 시퀀스 데이터는 (i) 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스, 및 (ii) 복수의 품질 점수들을 표현하며, 복수의 품질 점수들의 각각의 품질 점수는 판독 시퀀스의 특정 염기 호출이 핵산 시퀀싱 디바이스에 의해 올바르게 식별되었을 가능성을 나타냄 -, 하나 이상의 컴퓨터들에 의해, 복수의 품질 점수들의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 단계 - 각각의 품질 점수 그룹은 복수의 품질 점수들의 품질 점수들의 서브세트를 포함함 -, 복수의 품질 점수들의 제1 서브세트 내의 각각의 특정 품질 점수에 대해: 하나 이상의 컴퓨터들에 의해, 복수의 품질 점수들의 제1 서브세트 내의 특정 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아니라고 결정하는 단계, 및 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 단일 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제1 데이터를 생성하는 단계 - 축소된 시퀀스 내의 단일 엔트리로서 사용될 제1 데이터는 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수를 표현함 -, 및 하나 이상의 컴퓨터들에 의해, 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 생성된 제1 데이터를 집계함으로써 축소된 시퀀스를 생성하는 단계를 포함할 수 있다.

[0025] 다른 버전들은 컴퓨터 판독가능 저장 디바이스들 상에 인코딩된 명령어들에 의해 정의되는 방법들의 액션들을 수행하기 위한 대응하는 시스템들, 장치, 및 컴퓨터 프로그램들을 포함한다.

[0026] 이들 및 다른 버전들은 선택적으로 하기의 특징들 중 하나 이상을 포함할 수 있다. 예를 들어, 일부 구현예들에서, 획득된 데이터는 FASTQ 파일을 포함한다.

[0027] 일부 구현예들에서, 복수의 품질 점수들 내의 각각의 품질 점수는 품질 점수에 대한 ASCII 값을 표현하는 데이터이다.

[0028] 일부 구현예들에서, 방법은, 복수의 품질 점수들의 제2 서브세트 내의 각각의 특정 품질 점수에 대해: 하나 이상의 컴퓨터들에 의해, 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의

멤버라고 결정하는 단계, 및 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버라는 결정에 기초하여, 하나 이상의 컴퓨터들에 의해 그리고 미리 결정된 그룹 매핑을 사용하여, 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터를 생성하는 단계를 추가로 포함할 수 있으며, 여기서 축소된 시퀀스 내의 단일 엔트리로서 사용될 제2 데이터는 특정 품질 점수 그룹을 표현하고, 미리 결정된 그룹 매핑은 복수의 상이한 품질 점수 그룹들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의한다.

[0029] 일부 구현예들에서, 하나 이상의 컴퓨터들에 의해, 축소된 시퀀스를 생성하는 단계는, 하나 이상의 컴퓨터들에 의해, 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버가 아닌 품질 점수들 각각에 대해 생성된 제1 데이터를 집계하는 단계, 및 하나 이상의 컴퓨터들에 의해, 특정 품질 점수 그룹들 각각에 대해 생성된 제2 데이터를 집계하는 단계를 포함할 수 있다.

[0030] 일부 구현예들에서, 본 방법은, 하나 이상의 컴퓨터들에 의해, 복수의 품질 점수들에서 복수의 품질 점수 그룹들을 식별하는 단계를 추가로 포함할 수 있다.

[0031] 일부 구현예들에서, 미리 결정된 단일 매핑은 복수의 단일 품질 점수들 각각과 대응하는 단일 엔트리 사이의 일대일 매핑을 정의한다.

[0032] 본 개시내용의 이들 및 다른 혁신적인 양태들은 상세한 설명, 도면 및 첨부된 청구범위를 참조하여 본 명세서의 아래에서 설명된다.

도면의 간단한 설명

[0033] 도 1은 제1 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩, 예를 들어, 압축하기 위한 시스템의 예를 도시하는 도면이다.

도 2는 제1 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩, 예를 들어, 압축하기 위한 프로세스의 예를 예시하는 흐름도이다.

도 3은 제2 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩, 예를 들어, 압축하기 위한 시스템의 예를 도시하는 도면이다.

도 4는 제2 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩, 예를 들어, 압축하기 위한 프로세스의 예를 예시하는 흐름도이다.

도 5는 제1 데이터 포맷을 갖는 품질 점수들의 시퀀스를 압축해제하기 위한 프로세스의 예를 예시하는 흐름도이다.

도 6은 제2 데이터 포맷을 갖는 품질 점수들의 시퀀스를 압축해제하기 위한 프로세스의 예를 예시하는 흐름도이다.

도 7은 품질 점수들의 압축 방법을 결정하기 위한 프로세스의 예를 예시하는 흐름도이다.

도 8은 제1 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩하기 위한 프로세스의 실험 결과들의 그래픽 묘사이다.

도 9는 제2 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩하기 위한 프로세스의 실험 결과들의 그래픽 묘사이다.

도 10은 제1 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩하기 위한 프로세스를 구현하는 데 사용될 수 있는 컴퓨터 시스템 컴포넌트들의 도면이다.

다양한 도면들에서 동일한 도면 부호들 및 명칭들은 동일한 요소들을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0034] 본 개시내용은 핵산 시퀀싱 디바이스에 의해 생성된 판독 시퀀스들에 대한 품질 점수들의 시퀀스를 표현하는 데이터의 압축을 위한 방법, 시스템 및 컴퓨터 프로그램에 관한 것이다. 품질 점수들의 시퀀스 내의 각각의 품질 점수는 판독 시퀀스 내의 대응하는 염기가 핵산 시퀀싱 디바이스에 의해 올바르게 시퀀싱되었을 가능성의 표시를 제공한다. 본 명세서에 개시된 방법들 및 시스템들은 본 명세서에 설명된 기법들을 이용하지 않는 종래의 방법들과 비교할 때 더 빠른 압축 속도 및 더 작은 압축비를 가능하게 한다. 더 빠른 압축 속도 및 더 작은 압축

비는, 압축 엔진에 의해 처리되는 품질 점수들의 시퀀스들을 표현하는 입력 데이터의 크기를 감소시키기 위해 사전-압축 인코딩 단계들을 수행함으로써 달성된다. 압축 엔진은 품질 점수들의 시퀀스들을 표현하는 감소된 크기의 입력을 수신하고 처리하기 때문에, 압축 엔진은 종래의 방법들에 비해 입력 데이터의 압축을 더 빠르게 수행하고 크기가 더 작은 압축된 파일들을 달성할 수 있다. 따라서, 본 개시내용의 압축 방법들은 종래의 시스템들보다 더 높은 압축비를 달성할 수 있으며, 여기서 압축비는 비압축된 파일 크기를 압축된 파일 크기로 나눈 것과 동일하다.

[0035] 일반적으로, 본 개시내용은 품질 점수 시퀀스를 표현하는 데이터를 압축 엔진에 입력하기 전에 품질 점수 시퀀스를 표현하는 데이터에 대해 초기 인코딩 단계를 실행하는 시스템들 및 방법들을 설명한다. 이 접근법의 이득은 특정 예와 관련하여 설명될 수 있다. 하나 이상의 제1 구현예들에서, 품질 점수들의 시퀀스 내의 각각의 품질 점수는 8비트(또는 1바이트) ASCII 값으로 표현될 수 있다. 본 명세서에 설명된 그러한 제1 구현예들에서, 본 개시내용은 품질 점수들의 시퀀스 내의 품질 점수의 그러한 8비트 표현에 대해 초기 인코딩을 수행하여 품질 점수의 8비트 표현을 품질 점수의 2비트 표현 또는 1.6비트 표현으로 감소시킬 수 있으며, 그에 따라 4개의 품질 점수들 또는 5개의 품질 점수들이 각각 단일 바이트로 인코딩되게 할 수 있다. 따라서, 이러한 초기 인코딩 단계는 압축 엔진에 대한 입력 데이터 크기를 품질 점수들의 초기 표현의 크기의 1/4 또는 1/5만큼 상당히 그리고 예측가능하게 감소시킬 수 있다.

[0036] 그러나, 본 개시내용은 압축 엔진에 대한 입력을 위해 품질 점수의 8비트 표현을 품질 점수의 2비트 표현 또는 품질 점수의 1.6비트 표현으로 감소시키는 것으로 제한되지 않는다. 오히려, 품질 점수들의 다른 크기의 표현들로 유사한 비율 감소가 달성될 수 있다. 이들 예는, 부분적으로, 본 개시내용에 의해 달성되는 기술적 개선들을 강조하기 위해 제공되고 있다.

[0037] 더욱이, 본 개시내용의 다른 제2 구현예들은 압축 엔진으로의 입력을 위한 축소된 시퀀스 세트를 생성하기 위해 품질 점수들의 시퀀스를 표현하는 데이터에 대한 동작들을 수행하는 다른 초기 단계 인코딩 엔진들을 설명한다. 이러한 제2 구현예들은 위에서 설명된 구현예들과 유사한 기술적 이득들을 제공한다(예를 들어, 종래의 방법들에 비해 더 빠른 압축 속도 및 더 높은 압축비). 그러나, 품질 점수들의 자발적인 그룹화 또는 비-그룹화와 관련되고 가변적인, 압축 엔진에 입력하기 전의 입력 데이터의 초기 크기 감소로 인해, 이러한 제2 구현예들은, 비록 여전히 종래의 방법들보다 더 빠른 압축 속도, 더 낮은 압축된 파일 크기, 및 더 높은 압축비를 생성하지만, 궁극적으로는, 각각의 단계에서 정의된 데이터 크기들을 가질 수 있는 위에서 설명된 본 개시내용의 제1 구현예들보다 덜 예측가능한, 더 빠른 속도 및 압축비를 가질 수 있다.

[0038] 일반적으로, 본 명세서에서 사용되는 용어 "인코딩"은, 제1 세트의 데이터를 수신하는 것 및 제1 세트의 데이터를 처리하여 제1 세트의 데이터를 상이한 형태로 표현하는 제2 세트의 데이터를 생성하는 것을 포함하는, 하나 이상의 소프트웨어 엔진들, 하나 이상의 하드웨어 엔진들(예를 들어, 프로세서들), 또는 이들의 조합에 의해 수행되는 프로세스(들)를 지칭한다. 일부 실시예들에서, 제2 세트의 데이터는 수신된 제1 세트의 데이터보다 작은 메모리에 저장될 수 있다. 예를 들어, 데이터를 인코딩하는 하나의 형태는, 예를 들어 압축 엔진을 이용해, 압축 이전의 데이터의 크기보다 작은 크기로 데이터를 압축하는 것을 포함할 수 있다.

[0039] 도 1은 제1 데이터 포맷을 갖는 품질 점수들의 시퀀스를 압축하기 위한 시스템(100)의 예를 도시하는 도면이다. 제1 데이터 포맷은 "X"개의 상이한 품질 점수들을 판독 시퀀스 내의 대응하는 염기에 할당할 수 있으며, 여기서 "X"는 주어진 임계치보다 작은 임의의 양의(positive) 정수이다. 주어진 임계치는 도 3에 도시된 시스템(300)과 같은 다른 시스템들보다 더 효과적이거나 실용적인 도 1에 도시된 시스템(100)을 사용하여 생성하는 고유 품질 점수들의 개수에 기초하여 결정될 수 있다. 예를 들어, 주어진 임계치는 8일 수 있다. "X"가 8 미만인 경우, 대응하는 품질 점수들은 시스템(100)에 의해 처리될 수 있다. 특정 예로서, 일부 구현예들에서 "X"는 4와 같을 수 있으며, 이는 제1 데이터 포맷이, 핵산 시퀀싱 디바이스에 의해 생성된 판독 시퀀스의 하나 이상의 염기 호출들이 정확할 가능성을 나타내기 위해 4개의 상이한 품질 점수들 중 임의의 하나를 사용할 수 있음을 나타낸다. 일부 구현예들에서, 가능성은, 품질 점수에 대응하는 하나 이상의 염기 호출들에서 시퀀싱 오류가 발생했을 확률을 포함할 수 있다. 일부 실시예들에서, 시퀀싱 오류는 부정확한 판독 시퀀스 내의 특정 위치에 대해 핵산 시퀀싱 디바이스에 의해 생성된 염기 호출을 포함할 수 있다. 예를 들어, 시퀀싱 디바이스는, 사실상 정확한 염기 호출이 문자 C로 표현되는 시토신이었어야 할 때, 판독 시퀀스 내의 특정 위치에 대해 문자 A로 표현되는 아데닌의 염기 호출을 결정할 수 있다. 주어진 염기 호출에 대한 낮은 품질 점수는 그러한 오류가 가능성이 더 높다는 것을 나타낼 수 있는 반면, 높은 품질 점수는 그러한 오류가 가능성이 더 낮다는 것을 나타낼 수 있다. 염기 호출은 판독 시퀀스의 특정 뉴클레오티드를 표현하는, 핵산 시퀀싱 디바이스에 의해 생성된 데이터를 포함할 수

있다.

[0040] 시스템(100)은 하나 이상의 데이터 소스들로부터 입력 데이터(102)를 수신하도록 구성된다. 일부 구현예들에서, 하나 이상의 데이터 소스들은 핵산 시퀀싱 디바이스를 포함할 수 있다. 핵산 시퀀싱 디바이스는 노바식(Novaseq)® 6000, 넥스트식(Nextseq)® 2000 등과 같은 차세대 시퀀싱 디바이스일 수 있다. 다른 구현예들에서, 하나 이상의 데이터 소스들은 태블릿 컴퓨터, 데스크톱 컴퓨터, 하나 이상의 서버 컴퓨터들, 또는 이들의 조합과 같은 컴퓨팅 디바이스들 상에서 동작하는 하나 이상의 프로세서들을 포함할 수 있다. 일부 구현예들에서, 입력 데이터(102)는 하나 이상의 네트워크들을 통해 하나 이상의 데이터 소스들로부터 수신될 수 있다. 하나 이상의 네트워크들은 유선 이더넷 네트워크, 유선 광 네트워크, 무선 네트워크, LAN, WAN, Wi-Fi 네트워크, 셀룰러 네트워크, 인터넷, 또는 이들의 임의의 조합을 포함할 수 있다. 일부 구현예들에서, 입력 데이터(102)는 USB 케이블 접속, USB-C 케이블 접속 등과 같은 직접 접속을 통해 하나 이상의 데이터 소스들로부터 수신될 수 있다. 또 다른 구현예들에서, 전체 시스템(100)은 하나 이상의 데이터 소스들 내에서 호스팅될 수 있다. 예를 들어, 일부 구현예들에서, 전체 시스템(100)은 핵산 시퀀싱 디바이스에 의해 호스팅될 수 있다.

[0041] 시스템(100)은 입력 데이터(102)를 수신하는 입력 엔진을 포함할 수 있다. 입력 데이터(102)는 일련의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스를 설명하는 데이터뿐만 아니라, 일련의 뉴클레오티드들 또는 염기들에 대한 품질 점수들을 설명하는 데이터를 각각 포함하는 복수의 레코드들을 포함할 수 있다. 염기 호출은 핵산 시퀀서 디바이스에 의해 시퀀싱된 DNA 단편의 뉴클레오티드에 대응하거나 이를 표현하는 핵산 시퀀서 디바이스에 의해 생성된 데이터를 포함할 수 있다. 그러나, 본 개시내용의 목적을 위해, 용어들 염기 호출 및 뉴클레오티드는 판독 시퀀스의 A, C, T 또는 G와 같은 문자에 대응하는 핵산 시퀀서에 의해 생성된 데이터를 지칭하기 위해 전반에 걸쳐 상호교환적으로 사용될 수 있다. 이러한 문자들의 의미는 아래에 더 상세히 설명되어 있다.

[0042] 각각의 레코드에서, 품질 점수들의 시퀀스 내의 각각의 품질 점수는 판독 시퀀스 내의 특정 뉴클레오티드 또는 염기에 대응할 수 있다. 예를 들어, 도 1의 예에서, 품질 점수들의 시퀀스 "F#FFFF ... F;FF"의 처음 품질 점수들 "F"는 판독 시퀀스 "CNTGTA ... ATAAG" 내의 첫 번째 뉴클레오티드 또는 염기에 대응한다. 일부 구현예들에서, 입력 데이터(102)는 하나 이상의 FASTQ 파일들을 포함할 수 있고, 복수의 레코드들의 각각의 레코드는 본 명세서에서 FASTQ 레코드로 지칭되는 FASTQ 파일들의 일부분을 포함할 수 있다. FASTQ 파일의 각각의 부분은 하나의 판독 시퀀스 및 그 판독 시퀀스에 대한 대응하는 품질 점수들의 시퀀스를 포함할 수 있다.

[0043] 각각의 판독 시퀀스는, 하나 이상의 핵산 시퀀서에 의해 생물학적 샘플을 시퀀싱하는 것으로부터, 하나 이상의 핵산 시퀀싱 디바이스들에 의해 이전에 생성될 수 있다. 생물학적 샘플은 인간, 동물 또는 식물과 같은 임의의 살아있는 유기체의 핵산 샘플을 포함할 수 있다. 각각의 판독 시퀀스는 정의된 어휘로부터의 문자열을 포함한다. 예를 들어, 가장 작은 어휘는 {A, C, G, T, N}의 5개의 심볼들의 세트에 의해 표현될 수 있다. 문자들 A, C, G 및 T는 데옥시리보핵산(DNA)에 존재하는 4개 유형의 뉴클레오티드들, 즉 아데닌, 시토신, 구아닌 및 티민을 표현한다. 리보핵산(RNA)에서, 티민은 우라실(U)로 대체된다. 문자 "N"은, 시퀀싱 디바이스가 "N"에 의해 점유된 판독 시퀀스의 특정 위치에서 어떠한 염기도 호출할 수 없었고 따라서 판독 시퀀스 내의 위치의 실제적이고 정확한 특성은 미결정됨을 나타내기 위해 핵산 시퀀싱 디바이스에 의해 사용될 수 있다. 문자들 A, C, G 및 T 또는 U의 사용은 일반적인데, 그 이유는 이러한 문자들이 각자의 뉴클레오티드들 각각의 첫 번째 문자를 표현하기 때문이다. 그러나, 본 개시내용은 미결정되는 생성된 판독 시퀀스 내의 위치를 표현하기 위해 문자 "N"을 사용하는 것으로 제한되지 않는다. 대신에, 임의의 문자 또는 심볼이, 핵산 시퀀싱 디바이스가 정확한 염기를 정확하게 호출할 수 없는 판독 시퀀스 내의 위치를 표현하기 위해 핵산 시퀀싱 디바이스에 의해 사용할 수 있다. 본 명세서에 기술된 구현예들에서, 미지의 염기를 표현하기 위해 사용된 문자 또는 심볼은 문자 "N"의 사용과 동등하다는 것이 이해된다.

[0044] 판독 시퀀스의 염기들에 대한 품질 점수들을 저장하는 것은 많은 유용한 응용들을 가질 수 있다. 그러나, 인간과 같은 살아있는 유기체에 대한 시퀀싱된 게놈이 주어지면, 30억 개 이상의 염기들 및 대응하는 30억 개 이상의 대응하는 품질 점수들을 포함할 수 있다. 결과적인 품질 점수들의 세트들에 대응하는 데이터는 매우 클 수 있고(예를 들어, 시퀀싱의 깊이에 따라, 수 기가바이트에서 테라바이트까지), 품질 점수 정보를 효율적으로 저장하거나 송신하거나 보관하기 위해 압축을 필요로 할 수 있다. 도 1을 참조하여 설명된 본 개시내용의 양태들은, 압축 엔진을 위한 입력 데이터 세트를 준비하기 위해 초기 인코딩 엔진에 의존하는 다단계 압축 방법을 제공한다. 압축 엔진의 추후 단계 사용 전에 입력 데이터(102)의 고유 속성들에 기초하여 데이터를 초기에 인코딩함으로써, 시스템(100)은 빠른 압축 속도를 유지하면서 높은 압축비를 달성할 수 있다. 결과적인 압축된 품질 점수들은, 결과적인 압축된 품질 점수들에 액세스하거나 이를 분석할 때 처리 속도를 증가시킬 뿐만 아니라 관

련 메모리 저장을 위한 운영 비용을 감소시킬 수 있는 감소된 메모리 풋프린트를 가질 수 있다. 본 개시내용의 목적을 위해, "엔진"은 하나 이상의 소프트웨어 모듈들, 하나 이상의 하드웨어 모듈들, 또는 이들의 임의의 조합을 포함할 수 있다.

[0045] 단계 A에서, 분류기 엔진(104)은 입력 데이터(102)를 획득하고, 입력 데이터(102)의 표적화된 또는 특정 부분들의 속성들에 기초하여 입력 데이터(102)의 특정 부분들을 상이한 초기 품질 점수 인코딩 엔진들로 라우팅할 수 있다. 본 개시내용의 맥락 내에서, 획득 프로세스는 수신, 검색, 또는 달리 획득하는 프로세스를 지칭한다. 일부 구현예들에서, 분류기 엔진(104)은 입력 데이터(102)의 각각의 특정 부분이 초기 품질 점수 인코딩 엔진 V1(106)에 제공되어야 하는지 또는 초기 품질 점수 인코딩 엔진 v2(118)에 제공되어야 하는지를 결정하는 결정 엔진으로서 기능할 수 있다. 그들이 수신하는 입력 데이터(102)의 부분들에 대해 그들 각자의 초기 인코딩 동작들을 수행한 후, 초기 인코딩 엔진들 V1, V2 각각은 압축 엔진(116)에 대한 입력으로서 그들 각자의 인코딩된 출력들(114, 126)을 제공할 수 있다. 압축 엔진은 그것이 입력들로서 수신하는 인코딩된 출력들(114, 126)을 처리할 수 있고, 입력 데이터(102)의 압축된 버전인 최종 출력(128)을 생성할 수 있다.

[0046] 도 1의 단계 A에서 보다 상세하게, 분류기 엔진(104)에 의해 획득된 입력 데이터(102)는 제1 레코드(102a), 제2 레코드(102b), 및 i번째 레코드(102c)를 포함할 수 있다. 도 1의 예에서, 제1 레코드(102a), 제2 레코드(102b), 및 i번째 레코드(102c)는 FASTQ 파일의 부분들일 수 있으며, 이는 본 명세서에서 FASTQ 레코드로도 지칭된다. FASTQ 레코드는 FASTQ 레코드가 유래된 FASTQ 파일을 식별하고 FASTQ 레코드를 다른 FASTQ 레코드들과 구별하는 레코드 헤더(예를 들어, "@A0:90:H46:1…")를 포함할 수 있다. FASTQ 레코드는 핵산 시퀀서에 의해 생성된 판독 시퀀스를 표현하는 데이터를 포함할 수 있다. FASTQ 레코드는 FASTQ 레코드의 시퀀스에 대응하는 품질 점수들의 시퀀스를 포함할 수 있다. FASTQ 레코드는 하나 이상의 데이터 구성요소들, 예를 들어, 레코드 헤더, 품질 점수 시퀀스, 시퀀스 등을 분리하기 위해 하나 이상의 구분자(delimiter)들을 추가로 포함할 수 있다.

[0047] 도 1의 예에서, 입력 데이터(102)는 3개의 데이터 레코드들을 갖는 것으로 도시되지만, 임의의 수의 레코드들이 입력 데이터(102) 내에 포함될 수 있다. 3개의 데이터 레코드들이 도 1의 후속 프로세스들에 나타나지만, 임의의 수의 레코드들이 유사한 방식으로 처리될 수 있다. 단계 B에서, 분류기 엔진(104)은, 하나 이상의 분류 규칙들에 기초하여, 입력 데이터(102)의 제1 레코드(102a), 제2 레코드(102b), 및 i번째 레코드(102c) 각각에 대한 분류들을 결정할 수 있다. 일부 구현예들에서, 분류들은 각자의 FASTQ 레코드들 내의 판독 시퀀스들의 염기 호출들에 기초하여 결정될 수 있다.

[0048] 일부 구현예들에서, 예를 들어, 분류기 엔진(104)은 FASTQ 레코드 내의 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하는지 여부에 기초하여 각각의 FASTQ 레코드를 초기 품질 점수 인코딩 엔진 V1, V2로 분류 또는 라우팅할 수 있다. FASTQ 레코드의 판독 시퀀스가 적어도 하나의 "N" 염기를 포함한다고 결정되는 경우, FASTQ 레코드는 초기 품질 점수 인코딩 엔진 V1(106)로 라우팅될 수 있다. 대안적으로, FASTQ 레코드의 판독 시퀀스가 적어도 하나의 "N" 염기를 포함하지 않는다고 결정되는 경우, FASTQ 레코드는 초기 품질 점수 인코딩 엔진 V2(118)로 라우팅될 수 있다. 그러나, 이것은 분류 규칙의 한 예일 뿐이며, 초기 품질 점수 인코딩 엔진들 V1, V2 사이에서 FASTQ 레코드를 분류하고 라우팅하기 위해 본 개시내용에 따라 다른 유형들의 분류기 규칙들이 사용될 수 있다고 간주된다. 일부 구현예들에서, 분류기 엔진(104)은 입력 데이터의 다른 요소들을 사용하여 입력 데이터를 분류한다. 예를 들어, "N" 염기들의 존재 또는 부재에 기초하여 분류하는 대신에, 분류기 엔진(104)은 "N" 염기들 또는 다른 염기들의 백분율 또는 결정된 부분에 기초하여 분류할 수 있다. 또한, 다른 염기 호출들 또는 대응하는 데이터, 예를 들어 품질 점수들은, 분류기 엔진(104)에 의해 하나 이상의 FAST Q 레코드들 또는 관련 데이터를 분류하거나 라우팅하는 데 사용될 수 있다.

[0049] 도 1의 예에서, 분류기 엔진(104)은, 단계 B에서 그리고 제1 FASTQ 레코드(102a) 및 i번째 FASTQ 레코드(102c)의 각자의 판독 시퀀스들의 염기 호출들에 기초하여, 제1 FASTQ 레코드(102a) 및 i번째 FASTQ 레코드(102c) 둘 모두가 적어도 하나의 "N" 염기를 갖는 판독 시퀀스를 표현하는 데이터를 포함한다고 결정할 수 있다. 제1 FASTQ 레코드(102a) 및 i번째 FASTQ 레코드(102c)가 각각 적어도 하나의 "N" 염기를 갖는다는 결정에 기초하여, 분류기 엔진(104)은 제1 FASTQ 레코드(102a) 및 i번째 FASTQ 레코드(102c)를 초기 품질 점수 인코딩 엔진 v1(106)로 라우팅할 수 있다. 일부 구현예들에서, 분류기 엔진(104)은, 예를 들어, 제1 FASTQ 레코드(102a) 또는 i번째 FASTQ 레코드(102c)에 대응하는 하나 이상의 완전한 FASTQ 레코드를 전송하지 않고, 제1 FASTQ 레코드(102a) 및 i번째 FASTQ 레코드(102c)에 대응하는 데이터를 라우팅한다. 제1 FASTQ 레코드(102a) 및 i번째 FASTQ 레코드(102C)에 대응하는 데이터를 초기 품질 점수 인코딩 엔진 v1(106)로 라우팅한 후, 시스템(100)의 실행은 단계 C에서 계속될 수 있다.

[0050] 단계 C에서, 초기 품질 점수 인코딩 엔진 v1(106)은 제1 FASTQ 레코드(102a) 및 i번째 FASTQ 레코드(102c)의 판독 시퀀스에 각각 대응하는 제1 FASTQ 레코드(102a) 및 i번째 FASTQ 레코드(102c)의 각자의 품질 점수 시퀀스들을 획득할 수 있다. 이 예에서, 초기 품질 점수 인코딩 엔진 v1(106)에 의해 수신된 품질 점수 시퀀스들은 4개의 품질 점수들을 포함할 수 있고, 4개의 품질 점수들 각각은 8비트 ASCII 값 "F", ":", "," 및 "#"로 표현되며, 이때 "#"은 "N" 염기들에 대한 품질 점수를 표현한다. 그러나, 다른 구현예들에서는, 유사한 정보를 나타내기 위해 사용될 수 있는 다른 문자들 또는 심볼들과 함께, 4개 미만 또는 그 초과의 품질 점수들이 있을 수 있다.

[0051] 초기 품질 점수 인코딩 엔진 v1(106)은, 단계 C에서, 제1 레코드(102a) 내의 품질 점수들의 시퀀스에 대한 초기 인코딩을 수행할 수 있다. 초기 품질 점수 인코딩 엔진 v1(106)에 의해 수행된 이러한 초기 인코딩은 제1 FASTQ 레코드(102a) 내의 품질 점수들의 시퀀스로부터 각각의 품질 점수를 인코딩할 수 있다. 이러한 초기 인코딩은 품질 점수들 "F", ":", "," 또는 "#" 각각의 8비트 ASCII 표현을 각각의 개별 품질 점수의 2비트 표현으로 인코딩하는 것을 포함할 수 있다. 일부 구현예들에서, 제1 FASTQ 레코드(102a)의 품질 점수 시퀀스의 각각의 품질 점수의 각각의 8비트 ASCII 표현의 인코딩은, 초기 품질 점수 인코딩 엔진(106)이 바이트당 4개의 품질 점수들을 포함하는 출력 데이터(114)를 생성하게 할 수 있다. 초기 품질 점수 인코딩 엔진 v1(106)에 의해 생성된 이러한 인코딩 비율은 압축 엔진(116)에 대한 입력 데이터 레코드들의 크기를 4배만큼 감소시킨다. 다른 구현예들에서, 더 많거나 더 적은 품질 점수들, 또는 상이한 데이터 크기들의 품질 점수들을 하나 이상의 정보 비트들로 압축함으로써 다른 압축비들이 실현될 수 있다.

[0052] 초기 품질 점수 인코딩 엔진 v1(106)은 초기 품질 점수 패킹 엔진 v1(106)으로 라우팅되는 각각의 FASTQ 레코드에 대해 초기 인코딩 프로세스를 계속 수행할 수 있다. 예를 들어, 초기 품질 점수 인코딩 엔진 v1(106)은 i번째 FASTQ 레코드(102c)를 수신하고, 품질 점수들의 시퀀스로부터의 품질 점수들을 품질 점수들의 8비트 ASCII 값으로부터 2비트 표현으로 인코딩할 수 있다. 이 프로세스는, 초기 품질 점수 패킹 엔진 v1에 의해 수신된 품질 점수들의 시퀀스가 각각 처리되어 그들의 초기 인코딩을 2비트 품질 점수들로 생성할 때까지, 계속될 수 있다.

[0053] 보다 상세하게는, 초기 품질 점수 인코딩 엔진 v1(106)은 항목들(108, 110, 112)에 도시된 프로세스를 실행함으로써 8비트 품질 점수들을 2비트 품질 점수들의 출력 이진 스트림(114)으로 인코딩하는 것을 수행할 수 있다. 품질 점수 인코딩 엔진 v1(106)은 제1 레코드(102a)로부터 처음 4개의 품질 점수들을 획득할 수 있다. 일부 구현예들에서, 이것은, 초기 품질 점수 인코딩 엔진 v1(106)이 품질 점수들의 각자의 ASCII 값을 획득하는 것을 포함할 수 있다. 다른 구현예들에서, 이것은 품질 점수들의 다른 표현들을 획득하고, 110에 도시된 바와 같이 각각의 개별 품질 점수를 대응하는 ASCII 값에 매핑하는 것을 포함할 수 있다. 예를 들어, 처음 4개의 품질 점수 "F", "#", "F" 및 "F"는 초기 품질 점수 인코딩 엔진 v1(106)에 의해 그들의 대응하는 ASCII 값들 70, 35, 70 및 70에 매핑될 수 있다. 'N'을 갖는 제1 데이터 포맷을 사용하는 다른 품질 점수들에 대한 매핑들이 표(110)에 도시되어 있다. 이어서, 품질 점수들의 ASCII 표현은 각자의 품질 점수 범주에 매핑될 수 있으며, 여기서 품질 점수의 각각의 범주에 대응하는 하나의 범주가 있다. 이 예에서, 품질 점수 범주들은, 4개의 가능한 품질 점수들의 범위가 있으므로, 0, 1, 2 및 3으로 표현된다. 그러나, 사용 중인 품질 점수 시스템에 기초하여 더 많은(예를 들어, 0, 1, 2, 3, 4, 5 등) 범주들 또는 더 적은 범주들(예를 들어, 0, 1)이 있을 수 있다. 이 예에서 0은 미지의 염기에 대응하고 3은 높은 염기 품질 점수에 대응하지만, 본 개시내용은 이 구현예에 제한되지 않는다. 오히려 그리고 예를 들어, 정수 대신에, 100 중 백분율, 언어 기반 점수 값들, 예를 들어, 낮음, 중간 및 높음, 또는 당업계에 공지된 다른 표시 값들이 품질 점수들을 표현하는 데 사용될 수 있다.

[0054] 도 1의 예에서, 초기 품질 점수 인코딩 엔진 v1(106)은 4개의 ASCII 기반 품질 점수들 'F', '#', 'F' 및 'F'에 대응하는 품질 점수 값들 3, 0, 3 및 3을 사용하여 계산을 수행할 수 있다. 초기 품질 점수 인코딩 엔진 v1(106)은 $q_1 + 4 * q_2 + 4^2 * q_3 + 4^3 * q_4$와 같은 공식을 사용하여 품질 인코딩된 점수를 계산할 수 있으며, 여기서 q_1, q_2, q_3, 및 q_4는 품질 점수 값들을 표현한다. 3, 0, 3, 및 3의 품질 점수 값들의 경우, 공식은 $3 + 4 * 0 + 4^2 * 3 + 4^3 * 3$로서 평가될 수 있으며, 이는 항목(112)에 도시된 바와 같이 243과 같다. 초기 품질 점수 인코딩 엔진 v1(106)은 수 243의 이진 표현을 생성함으로써 이들 4개의 품질 점수의 대응하는 이진 표현을 생성할 수 있다. 이 이진 표현은 11110011이며, 이는 이어서 출력 이진 스트림(114)에 추가될 수 있다. 이 프로세스는, 제1 FASTQ 레코드(102a)에 대한 품질 점수들의 시퀀스 내의 품질 점수들 각각이 초기에 출력 이진 스트

림(114)으로 인코딩될 때까지, 반복적으로 계속될 수 있다.

[0055] 제1 FASTQ 레코드(102a)의 품질 점수들의 시퀀스의 초기 인코딩 후, 초기 품질 점수 인코딩 엔진 v1(106)은 수신된 각각의 후속 FASTQ 레코드에 대해 초기 인코딩 프로세스를 계속 수행할 수 있다. 예를 들어, 초기 품질 점수 인코딩 엔진 v1(106)은 제1 FASTQ 레코드(102a)를 참조하여 위에서 설명된 것과 동일한 방식으로 i번째 FASTQ 레코드(102c)의 초기 인코딩을 계속 수행할 수 있다. 초기 품질 점수 인코딩 엔진 v1(106)은 계속해서 후속 품질 점수들을 획득하고, 항목(108)에 도시된 초기 4개의 품질 점수들과 유사한 방식으로 초기 인코딩을 수행할 수 있다.

[0056] 단계 D에서, 초기 품질 점수 인코딩 엔진 v1(106)은 출력 이진 스트림(114)을 압축 엔진(116)으로 전송할 수 있다. 이어서, 압축 엔진(116)은 미리 결정된 압축 방법에 대응하는 후속 압축을 출력 이진 스트림(114)에 대해 수행할 수 있다. 일반적으로, 출력 이진 스트림(114)의 크기를 추가로 압축하거나 압축하기 위해 압축 엔진(116)에 의해 임의의 압축 프로세스가 채용될 수 있다. 예를 들어, 일부 구현예들에서, 압축 엔진(116)은 ZSTD(Zstandard) 라이브러리의 레벨 11을 사용하여 압축을 수행할 수 있다. 그러나, 본 개시내용은 그렇게 제한되지 않는다. 대신에, 일부 구현예들에서, ZSTD 라이브러리의 다른 레벨들뿐만 아니라 다른 압축 라이브러리들을 포함하는 다른 압축 방법들이 사용될 수 있다. 일반적으로, 당업계에 공지된 임의의 압축 방법 또는 압축 방법들의 조합이 사용될 수 있다.

[0057] 다른 경우들에서, 분류기 엔진(104)은, 하나 이상의 분류 규칙들을 적용함으로써, 수신된 FASTQ 레코드가 적어도 하나의 "N"을 포함하지 않는 판독 시퀀스를 포함한다고 결정할 수 있다. 수신된 FASTQ가 적어도 하나의 "N"을 포함하지 않는다는 결정에 기초하여, 분류기 엔진(104)은 제2 FASTQ 레코드(102b)에 대응하는 데이터를 초기 품질 점수 인코딩 엔진 v2(118)로 라우팅할 수 있다.

[0058] 제2 FASTQ 레코드(102b)는 어떠한 N도 포함하지 않기 때문에, 잠재적 품질 점수들의 범위는 잠재적 품질 점수들의 개수에서 1을 뺀 수일 수 있다. 즉, 분류기 엔진(104)이 "X"개의 잠재적 품질 점수들을 포함하는 것으로 제2 FASTQ 레코드를 결정하는 경우, 초기 품질 점수 인코딩 엔진 v2(118)에서 처리될 수 있는 잠재적 품질 점수들은 "X" - 1이다. 이 구현예에서 품질 점수 시퀀스는 3개의 별개의 품질 점수들만을 갖는다. 따라서, 이 예에서는, 3개의 별개의 품질 점수들만이, 부분적으로, 더 높은 압축비를 실현하기 위해 별도로 압축되어야 한다.

[0059] 단계 E에서, 초기 품질 점수 인코딩 엔진 v2(118)는 제2 레코드(102b)의 품질 점수 시퀀스를 획득할 수 있다. 항목(120)은 제2 레코드(102b)의 품질 점수 시퀀스의 품질 점수들의 서브세트를 도시한다. 보다 상세하게는, 초기 품질 점수 인코딩 엔진 v2(1118)는 항목들(120, 122, 124)에 도시된 프로세스를 실행함으로써 8비트 품질 점수들을 1.6비트 품질 점수들의 출력 이진 스트림(126)으로 인코딩하는 것을 수행할 수 있다. 초기 품질 점수 인코딩 엔진 v2(118)는 제2 FASTQ 레코드(102b)로부터 결정된 개수의 품질 점수들, 예를 들어 5개의 품질 점수들을 획득할 수 있다. 일부 구현예들에서, 이것은, 초기 품질 점수 인코딩 엔진 v2(118)가 단지 품질 점수들의 각자의 ASCII 값을 획득하는 것을 포함할 수 있다. 다른 구현예들에서, 이것은 품질 점수들의 다른 표현을 획득하고, 122에 도시된 바와 같이 각각의 개별 품질 점수를 대응하는 ASCII 값에 매핑하는 것을 포함할 수 있다. 예를 들어, 처음 5개의 품질 점수 ":", "F", ":", "," 및 "F"는 초기 품질 점수 인코딩 엔진 v2(118)에 의해 그들의 대응하는 ASCII 값들 58, 70, 58, 44 및 70에 매핑될 수 있다.

[0060] 이어서, 품질 점수들의 ASCII 표현은 각자의 품질 점수 범주에 매핑될 수 있으며, 여기서 품질 점수의 각각의 범주에 대응하는 하나의 범주가 있다. 이 예에서, 품질 점수 범주들은 0, 1 및 2에 의해 표현되는데, 이는 3개의 가능한 품질 점수들의 범위가 있기 때문이다(즉, X개의 가능한 품질 점수들 - 1, 그 이유는 초기 품질 점수 인코딩 엔진 v2(118)로 라우팅되는 FASTQ 레코드에 "N" 염기가 없기 때문임). 이 예에서, 초기 품질 점수 인코딩 엔진 v2(118)는 8비트 ASCII 품질 점수를 기수-3 수로 표현함으로써 8비트 ASCII 품질 점수들을 1.6비트 품질 점수로 인코딩할 수 있다. 기수-3은, 여기에 품질 점수들의 3개의 고유 범주들이 있기 때문에 사용된다.

[0061] 그러나, 본 개시내용은 전술된 예에 제한되지 않는다. 대신에, 총 8개의 고유 품질 점수 범주들과 같은 더 많은 품질 점수 범주들이 있는 구현예들에서, 초기 품질 점수 인코딩 엔진 v2(118)는 7개의 고유 품질 점수 범주들을 갖는 시퀀스들을 획득할 수 있으며, 여기서 원래의 8개의 고유 품질 점수 범주들은 "N" 염기에 대한 품질 점수를 포함한다. 그러한 구현예에서, 초기 품질 점수 인코딩 엔진 v2(118)의 초기 인코딩들을 표현하기 위해 기수-7 수가 사용될 수 있다.

[0062] 항목(112)과 유사하게, 항목(124)은 품질 점수들의 제1 세트와 관련된 초기 인코딩 엔진(118)에 의해 수행된 계산을 나타낸다. 이러한 경우에서, 초기 품질 점수 패킹 엔진 v2(118)는 $q_1 + 3 * q_2 + 3^2 * q_3 + 3^3 * q_4 + 3^4 * q_5$

와 같은 공식을 사용하여 품질 패킹된 점수를 계산하며, 여기서 q_1, q_2, q_3, q_4, 및 q_5는 품질 점수 값들을 표현한다. 1, 2, 1, 0, 및 2의 품질 점수 값들의 경우, 공식은 $1+3*2+3^2*1+3^3*0+3^4*2$로서 평가될 수 있으며, 이는 항목(124)에 도시된 바와 같이 178과 같다. 이어서, 178의 이진 표현, 예를 들어, 10110010은 출력 이진 스트림(126)에 추가될 수 있다.

[0063] 도 1의 예에서, 초기 품질 점수 패킹 엔진 v2(118)는 5개의 8비트 품질 점수들을 1바이트의 단일 ASCII 문자로 인코딩함으로써, 제2 레코드(102b)의 각각의 품질 점수를 5배만큼 압축할 수 있다. 다른 구현예들에서, 더 많거나 더 적은 품질 점수들, 또는 상이한 데이터 크기들의 품질 점수들을 하나 이상의 정보 비트들로 압축함으로써 다른 압축비들이 실현될 수 있다.

[0064] 초기 품질 점수 인코딩 엔진 v1(106)과 유사하게, 초기 품질 점수 인코딩 엔진 v2(118)는 후속 품질 점수들을 획득하고, 항목(120)에 도시된 초기 5개의 품질 점수들과 유사한 방식으로 인코딩을 수행할 수 있다. 단계 F에서, 초기 품질 점수 인코딩 엔진 v2(118)는 압축 엔진(116)에 대한 입력으로서 출력 이진 스트림(126)을 제공할 수 있다. 이어서, 압축 엔진(116)은 미리 결정된 압축 방법에 대응하는 후속 압축을 출력 이진 스트림(126)에 대해 수행할 수 있다. 예를 들어, 일부 경우들에서, 압축 엔진(116)은 바이트 스트링을 압축하기 위해 범위 인코더의 PPMD(Prediction by Partial Matching) 구현에 기초하여 압축을 수행한다. 일부 구현예들에서, 당업계에 공지된 다른 압축 방법들이 PPMD 구현 대신에 또는 이에 추가하여 사용될 수 있다. 일반적으로, 압축 엔진(116)은 당업계에 공지된 임의의 압축 방법과 같은 임의의 압축 또는 압축 방법들의 조합을 사용할 수 있다.

[0065] 일부 구현예들에서, 범위 인코더에 공급된 출력 이진 스트림(126)의 각각의 심볼은 선행 점수들에서 최고 품질 점수 값들의 개수를 표현하는 4비트 컨텍스트에 따라 계산될 수 있다. 예를 들어, 4비트 컨텍스트는 30개의 선행 품질 점수들 내의 최고 품질 점수들의 총 수를 2로 나눈 것으로서 계산될 수 있다. 이러한 특정 4비트 컨텍스트 공식은 4비트 컨텍스트에 맞는 0 내지 15의 범위의 값을 산출할 수 있다. 4비트 컨텍스트 또는 다른 컨텍스트 기반 인코딩 접근법들을 사용함으로써, 시스템(100)은 특정 품질 점수, 예를 들어 최고 점수 'F'의 다수의 인접한 인스턴스들을 고려할 수 있으며, 이는 후속 점수의 양호한 예측자일 수 있다. 일부 구현예들에서, 메모리, 계산 또는 다른 요건들에 따라 더 크거나 더 작은 컨텍스트들이 사용된다. 예를 들어, 더 큰 컨텍스트는 더 높은 압축비를 산출할 수 있지만, 더 많은 메모리 및 계산 시간을 필요로 할 수 있다.

[0066] 일부 구현예들에서, 다른 인코딩 엔진들이 시스템(100) 내에서 사용될 수 있다. 예를 들어, 초기 품질 점수 인코딩 엔진 v1(106) 및 초기 품질 점수 인코딩 엔진 v2(118)에 더하여, 구현은 분류기 엔진(104)에 의해 정의된 특정 기준에 기초한 제3 인코딩 엔진을 포함할 수 있다. 일부 구현예들에서, 3개 초과의 엔진들이 사용될 수 있다. 일부 구현예들에서, 분류기 엔진(104)은 2개의 별개의 품질 점수들만을 사용하는 입력 데이터(102) 내의 제3 세트의 레코드들을 결정할 수 있다. 이러한 방식으로, 시스템(100)은 2개의 별개의 품질 점수들만을 갖는 그룹에 대해 훨씬 더 큰 압축비들을 실현할 수 있다.

[0067] 설명의 편의를 위해 단계 A 내지 단계 G가 사용된다. 시스템(100)에 의해 수행되는 프로세스는 단계 A 내지 단계 G에 의해 표시된 순서로 발생할 수 있지만, 다른 구현예들에서, 특정 단계들의 순서는 상이할 수 있다. 일부 구현예들에서, 2개 이상의 단계가 동시에 발생할 수 있다.

[0068] 도 2는 제1 입력 데이터 포맷에 기초한 품질 점수 압축을 위한 프로세스(200)의 예를 예시하는 흐름도이다. 프로세스(200)는 하나 이상의 전자 시스템들, 예를 들어 도 1의 시스템(100)에 의해 수행될 수 있다.

[0069] 프로세스(200)는 유전자 시퀀싱 디바이스에 의해 생성된 유전자 판독 시퀀스를 획득하는 것(202)을 포함한다. 예를 들어, 도 1에 도시된 바와 같이, 입력 데이터(102)가 획득되어 시스템(100)에 제공된다. 입력 데이터(102)는 핵산 시퀀싱 디바이스에 의해 생성된 복수의 염기 호출들에 대응하는 데이터를 포함하는 판독 시퀀스를 표현하는 데이터뿐만 아니라, 복수의 염기 호출들에 대한 품질 점수들을 설명하는 데이터를 각각 포함하는 하나 이상의 레코드들을 포함할 수 있다. 품질 점수들의 각각의 품질 점수는 판독 시퀀스의 특정 염기 호출에 대응한다.

[0070] 일부 구현예들에서, 문자들은 입력 데이터(102) 내의 품질 점수들을 표현하기 위해 사용된다. 예를 들어, 제1 FASTQ 레코드(102a)의 품질 점수들의 시퀀스 "F#FFFF ... F;FF"의 처음 품질 점수들 "F"는 제1 FASTQ 레코드(102a)의 판독 시퀀스 "CNTGTA ... ATAAG" 내의 첫 번째 뉴클레오티드 또는 염기(이하에서 "염기"로 지칭됨)에 대응한다. 일부 구현예들에서, 입력 데이터(102)는 하나 이상의 FASTQ 파일들을 포함할 수 있고, 복수의 레코드들의 각각의 레코드는 본 명세서에서 FASTQ 레코드로 지칭되는 FASTQ 파일들의 일부분을 포함할 수 있다. FASTQ

파일의 각각의 부분은 하나의 판독 시퀀스 및 그 판독 시퀀스에 대한 대응하는 품질 점수들의 시퀀스를 포함할 수 있다.

[0071] 프로세스(200)는 유전자 판독 시퀀스에 대응하는 복수의 품질 점수들을 획득하는 것(204)을 포함할 수 있다. 예를 들어, 도 1에 도시된 바와 같이, 입력 데이터(102)는 제1 레코드(102a), 제2 레코드(102b) 및 i번째 레코드(102c)를 포함한다. 제1 레코드(102a), 제2 레코드(102b) 및 i번째 레코드(102c) 각각은 유전자 시퀀스 및 유전자 시퀀스에 대응하는 품질 점수 시퀀스 둘 모두를 포함한다. 예를 들어, 제1 레코드(102a)는 유전자 시퀀스 "CNTGTA...ATAAG" 및 대응하는 품질 점수 시퀀스 "F#FFFF...F:,FF"를 포함하며, 여기서 품질 점수 시퀀스의 각각의 값은 대응하는 유전자 시퀀스의 특정 위치에서의 시퀀싱 오류의 가능성을 나타낸다.

[0072] 프로세스(200)는 유전자 판독 시퀀스가 적어도 하나의 "N" 염기를 포함한다고 결정하는 것(206)을 포함한다. 예를 들어, 도 1에 도시된 바와 같이, 분류기 엔진(104)은 하나 이상의 유전자 판독 시퀀스들을 포함하는 입력 데이터(102)를 획득한다. 입력 데이터(102)는 제1 레코드(102a)를 포함한다. 제1 레코드(102a)는 유전자 시퀀스 "CNTGTA...ATAAG"를 포함한다. 유전자 시퀀스 "CNTGTA...ATAAG"는 염기 "N"을 포함한다. 분류기 엔진(104)은 제1 레코드(102a)의 유전자 시퀀스 "CNTGTA...ATAAG"가 염기 "N"을 포함한다고 결정할 수 있고, 제1 레코드(102a)에 대응하는 데이터를 초기 품질 점수 인코딩 엔진 v1(106)로 라우팅할 수 있다. 유사하게, 분류기 엔진(104)은 제2 레코드(102b)를 획득할 수 있다. 제2 레코드(102b)는 염기 "N"을 포함하지 않는 유전자 시퀀스 "GTCTAG...CACTT"를 포함한다. 분류기 엔진(104)은 제2 레코드(102b)의 유전자 시퀀스 "GTCTAG...CACTT"가 염기 "N"을 포함하지 않는다고 결정할 수 있고, 제2 레코드(102b)에 대응하는 데이터를 초기 품질 점수 인코딩 엔진 v2(118)로 라우팅할 수 있다.

[0073] 프로세스(200)는 기수-x 수를 사용하여 품질 점수들 각각을 인코딩함으로써 제1 인코딩된 데이터 세트를 생성하는 것을 포함하며, 여기서 x는 유전자 시퀀싱 디바이스(208)에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수이다. 예를 들어, 초기 품질 점수 인코딩 엔진 v1(106)은 제1 레코드(102a)의 품질 점수 시퀀스를 획득한다. 제1 레코드(102a)의 품질 점수 시퀀스는 4개의 고유 품질 점수들: "F", ":", "," 및 "#"를 포함한다. 다른 구현예들에서 다른 적합한 심볼들 또는 값들이 사용될 수 있다. 이어서, 초기 품질 점수 인코딩 엔진 v1(106)은 기수-4 수에 기초하여 정수를 계산할 수 있다. 예를 들어, 도 1의 항목(108)에 도시된 바와 같이, 초기 품질 점수 인코딩 엔진 v1(106)은 항목(110)에 도시된 매핑에 기초하여 유전자 품질 점수들 "F#FF"에 대응하는 값 3033을 생성한다. 이어서, 초기 품질 점수 인코딩 엔진 v1(106)은, 값 3033이 기수-4 표기의 형태로 기록된 것처럼 값 3033에 기초한 정수를 생성한다. 항목(112)에 도시된 바와 같이, 결과적인 정수는 243과 같으며, 이는 "11110011"로서 8개의 이진 비트들로 기록될 수 있다. 품질 점수들 "F#FF"와 같은 품질 점수들의 그룹의 이진 형태를 생성하는 데 사용되는 방정식은, 정수 값이 8개 이하의 비트들을 사용하여 표현될 수 있도록 구성될 수 있다. 예를 들어, 초기 품질 점수 인코딩 엔진 v1(106)에 의해 사용되는 방정식은 정수 값이 255 미만이 되도록 구성될 수 있다.

[0074] 예를 들어, 초기 품질 점수 인코딩 엔진 v2(118)는 제2 레코드(102b)의 품질 점수 시퀀스를 획득한다. 제2 레코드(102b)의 품질 점수 시퀀스는, 제1 레코드(102a)의 품질 점수 시퀀스와 대조적으로, 3개의 고유 품질 점수들 "F", ":" 및 ","를 포함한다. 다른 구현예들에서 다른 적합한 심볼들 또는 값들이 사용될 수 있다. 이어서, 초기 품질 점수 인코딩 엔진 v2(118)는 기수-3 수에 기초하여 정수를 계산할 수 있다. 더 적은 고유 품질 점수들이 있기 때문에, 초기 품질 점수 인코딩 엔진 v2(118)는 추가 품질 점수, 예를 들어, 4 대신 5개를, 8비트의 이진법으로 인코딩할 수 있다. 예를 들어, 도 1의 항목(124)에 도시된 바와 같이, 초기 품질 점수 인코딩 엔진 v2(118)는 항목(122)에 도시된 매핑에 기초하여 유전자 품질 점수들 ":F:,F"에 대응하는 값 12102를 생성한다. 이어서, 초기 품질 점수 인코딩 엔진 v2(118)는, 값 12102가 기수-3 표기의 형태로 기록된 것처럼 값 12102에 기초한 정수를 생성한다. 항목(124)에 도시된 바와 같이, 결과적인 정수는 178과 같으며, 이는 "10110010"으로서 8개의 이진 비트들로 기록될 수 있다. 품질 점수들 ":F:,F"와 같은 품질 점수들의 그룹의 이진 형태를 생성하는 데 사용되는 방정식은, 정수 값이 8개 이하의 비트들을 사용하여 표현될 수 있도록 구성될 수 있다. 예를 들어, 초기 품질 점수 인코딩 엔진 v2(118)에 의해 사용되는 방정식은 정수 값이 255 미만이 되도록 구성될 수 있다.

[0075] 프로세스(200)는 제2 인코딩 알고리즘을 사용하여 제1 인코딩된 데이터 세트를 인코딩함으로써 제2 인코딩된 데이터 세트를 생성하는 것(210)을 포함한다. 예를 들어, 도 1에 도시된 바와 같이, 압축 엔진(116)은 초기 품질 점수 인코딩 엔진 v1(106) 또는 초기 품질 점수 인코딩 엔진 v2(118)에 의해 제공된 입력에 기초하여 출력(128)을 생성한다. 일부 구현예들에서, 압축 엔진(116)은 상이한 압축 프로세스들로부터의 다수의 출력들을 조합하여 출력(128)을 생성한다. 예를 들어, 초기 품질 점수 인코딩 엔진 v1(106) 및 초기 품질 점수 인코딩 엔진

v2(118) 둘 모두는 인코딩된 출력들(114, 126)에 도시된 바와 같은 압축 엔진(116)에 대한 데이터를 생성할 수 있다.

[0076] 일부 구현예들에서, 압축 엔진(116)은 획득된 데이터에 기초하여 하나 이상의 유형의 압축을 수행한다. 예를 들어, 초기 품질 점수 인코딩 엔진 v1(106)의 출력을 압축하는 데 사용되는 압축 프로세스는 초기 품질 점수 인코딩 엔진 v2(118)의 출력을 압축하는 데 사용되는 압축 프로세스와 상이할 수 있다. 본 명세서에서 논의된 바와 같이, 초기 품질 점수 인코딩 엔진 v1(106)의 인코딩된 출력(114)의 압축은 ZSTD(Zstandard) 라이브러리의 레벨 11을 사용하는 압축 또는 다른 형태의 압축을 포함할 수 있다. 초기 품질 점수 인코딩 엔진 v2(118)의 인코딩된 출력(126)의 압축은 범위 인코더의 PPMD(Prediction by Partial Matching) 구현을 사용하는 압축 또는 다른 형태의 압축을 포함할 수 있다. 일부 구현예들에서, 시퀀스의 주어진 값 주위의 컨텍스트들은 주어진 시퀀스를 압축하는 데 사용될 수 있다. 예를 들어, 선행 점수들 내의 최고 품질 점수 값들의 개수를 표현하는 4비트 컨텍스트가 압축 엔진(116)에 의해 사용될 수 있다. 일부 경우들에서, 압축된 출력을 생성하기 위해 다수의 압축 프로세스들이 조합될 수 있다.

[0077] 도 3은 제2 데이터 포맷을 갖는 품질 점수들의 시퀀스를 압축하기 위한 시스템의 예를 도시하는 도면이다. 시스템(300)은 Q40 데이터 포맷에 기초하여 포맷화되는 품질 점수 시퀀스(302)를 포함한다. 품질 점수 시퀀스(302)는, 축소된 시퀀스(332) 및 품질 점수 시퀀스(302)의 압축된 버전을 표현하는 후속 출력(336)을 생성하기 위해, 그룹 식별 엔진(304), 빈도 카운터 엔진(310), 축소된 시퀀스 생성 엔진(316), 단일 매핑 엔진(320), 그룹 매핑 엔진(326), 및 압축 엔진(334)에 의해 처리된다.

[0078] 도 3의 단계 A에서, 품질 점수 시퀀스(302)가 생성되어 그룹 식별 엔진(304)으로 전송된다. 도 3의 예에서, 품질 점수 시퀀스(302)는 위에서 논의된 바와 같이 염기들과 관련된 품질 점수들을 인코딩하기 위해 Q40 데이터 포맷을 사용하는 시퀀서에 의해 생성된다. 일반적으로, 임계 수 초과의 고유 품질 점수들을 사용하는 임의의 시퀀서는 도 3에 도시된 시스템(300)에 의해 수행되는 프로세스 또는 유사한 프로세스를 사용하여 결과적인 출력 데이터를 압축할 수 있다. 예를 들어, 시퀀서가 유전자 시퀀스에 대응하는 품질 점수들을 인코딩하기 위해 8개 초과의 고유 품질 점수들을 사용하는 경우, 시퀀서는 도 3을 참조하여 논의된 인코딩 및 압축 프로세스들을 사용할 수 있다.

[0079] 단계 B에서, 그룹 식별 엔진(304)은 품질 점수 시퀀스(302)를 획득하고, 항목(306)에 도시된 바와 같이 품질 점수 시퀀스(302)에 기초하여 하나 이상의 그룹들을 생성한다. 이 구현예에서, 그룹 식별 엔진(304)은 인접한 품질 점수들을 3개의 그룹들로 그룹화한다. 다른 구현예들에서, 다른 개수의 품질 점수들이 하나 이상의 그룹들 내에 포함될 수 있다. 항목(316)에 도시된 바와 같이, 3개의 품질 점수들의 제1 그룹은 품질 점수 시퀀스(302)의 문자들에 대응하는 문자들 '@', 'C' 및 '@'이다. 도 3의 예에서, 품질 점수 시퀀스(302)의 각각의 문자는 시퀀싱 오류의 가능성을 나타내는 품질 점수를 표현한다.

[0080] 그룹 식별 엔진(304)은 품질 점수 그룹들(308)을 생성하고 품질 점수 그룹들(308)을 빈도 카운터 엔진(310)으로 전송한다. 단계 C에서, 빈도 카운터 엔진(310)은 품질 점수 그룹들(308)을 획득하고, 항목(312)에 도시된 바와 같이 품질 점수 그룹들(308) 내의 그룹들 각각에 대한 발생 횟수를 결정한다.

[0081] 일부 구현예들에서, 그룹 식별 엔진(304) 및 빈도 카운터 엔진(310)은, 적어도 부분적으로, 병렬로 작동한다. 예를 들어, 그룹 식별 엔진(304)은 품질 점수 시퀀스(302)에 기초하여 단일 그룹을 식별할 수 있다. 이어서, 그룹 식별 엔진(304)은 식별된 단일 그룹을 빈도 카운터 엔진(310)으로 전송할 수 있다. 이어서, 빈도 카운터 엔진(310)은 품질 점수들의 식별된 단일 그룹에 대한 발생 횟수를 결정할 수 있다. 일부 경우들에서, 품질 점수 시퀀스(302)로부터 그룹을 생성하는 것은 품질 점수 시퀀스(302)의 하나 이상의 품질 점수들을 식별하는 것을 포함할 수 있다. 일반적으로, 본 명세서에 설명된 임의의 프로세스는 다른 프로세스와 동시에 스레드되거나 실행될 수 있으며, 2개 이상의 프로세스들은 하나 이상의 디바이스들 또는 소프트웨어의 인스턴스들 상에서 실행될 수 있다.

[0082] 빈도 카운터 엔진(310)은 품질 점수 그룹 빈도들(314)을 생성하고, 품질 점수 그룹 빈도들(314)을 축소된 시퀀스 생성 엔진(316)으로 전송한다. 품질 점수 그룹 빈도(314)는 품질 점수 시퀀스(302) 내의 하나 이상의 품질 점수들의 발생 횟수를 포함할 수 있다. 단계 D에서, 축소된 시퀀스 생성 엔진(316)은 품질 점수 그룹 빈도들(314)을 획득하고 단일 매핑 엔진(320) 및 그룹 매핑 엔진(326) 둘 모두와 통신할 수 있다. 즉, 축소된 시퀀스 생성 엔진(316)은, 품질 점수 시퀀스(320)의 품질 점수가 임계 수의 품질 점수들보다 많은 품질 점수들을 갖는 품질 점수들의 그룹의 일부가 아닌 경우 축소된 시퀀스(332) 내의 엔트리를 생성하기 위해 단계 E에서 단일 매핑 엔진(320)을 사용할 수 있다. 대안적으로, 축소된 시퀀스 생성 엔진(316)은, 품질 점수 시퀀스(320)의 품질

점수가 임계 수의 품질 점수들보다 많은 품질 점수들을 갖는 품질 점수들의 그룹의 일부인 경우 단계 F에서 그룹 매핑 엔진(326)을 사용할 수 있다. 본 개시내용의 목적을 위해, 축소된 시퀀스(332)의 "엔트리" 또는 "단일 엔트리"는 단일 품질 점수 또는 품질 점수들의 그룹을 대체하기 위해 사용된 축소된 시퀀스(332) 내의 "72"와 같은 단일 값을 포함할 수 있다.

[0083] 본 명세서의 목적을 위해, 품질 점수들의 임계 수의 양수 또는 음수 표현들 및 그룹 내의 품질 점수들의 개수를 사용함으로써 품질 점수들의 그룹이 품질 점수들의 임계 수보다 많은 품질 점수들을 갖는지 여부를 결정할 수 있다. 따라서, 품질 점수들의 그룹 내의 품질 점수들의 개수가 임계치를 "충족"하는지 여부 - 단지 품질 점수들의 개수가 임계치보다 많은지 여부가 아님 - 를 결정하는 것은 본 명세서와 일치한다. 이는, 이러한 관계가 3개의 임계치의 품질 점수들을 초과하는 4개의 품질 점수들을 갖는 그룹으로서 설명될 수 있거나 또는 음의 4개의 품질 점수들을 갖는 품질 점수들의 그룹이 음의 3개의 임계치의 품질 점수들보다 크지 않기 때문이다. 어느 경우든, 품질 점수들의 그룹은 임계치가 어떻게 구현되는지에 관계없이 3개 초과의 품질 점수들을 갖는다.

[0084] 설명의 편의를 위해, 본 명세서는 이제 그룹 매핑 엔진(326)의 프로세스들을 설명하기 전에 단일 매핑 엔진(320)의 프로세스들을 설명한다. 그러나, 그룹 매핑 엔진(326)이 그룹 매핑 프로세스들을 완료하기 전에 단일 매핑 엔진(320)이 단일 매핑 프로세스들을 완료할 필요가 없도록, 동시 처리 및 다른 유사한 방법들이 사용될 수 있다. 대신에, 축소된 시퀀스 생성 엔진(316)이 단일 매핑 엔진(320)을 호출하는지 또는 그룹 매핑 엔진(326)을 호출하는지 여부는, 축소된 시퀀스 생성 엔진(326)에 의해 처리되는 품질 점수들의 시퀀스(302)의 특정 품질 점수에 기초하여 결정될 것이다.

[0085] 단계 E에서, 단일 매핑 엔진(320)은 품질 점수 시퀀스(302)를 획득한다. 단일 매핑 엔진(320)은 항목(321)에 도시된 바와 같이 품질 점수 시퀀스(302) 및 단일 매핑 문자 목록(322)을 사용하여 미리 결정된 단일 매핑(324)을 생성한다. 도 3의 예에서, 단일 매핑 문자 목록(322)은 0 내지 63의 정수 값들을 포함한다. 품질 점수 시퀀스(302)의 ASCII 값들은 33 내지 96의 값들을 포함한다. 이러한 방식으로, 품질 점수 시퀀스(302)의 각각의 값은 0 내지 63의 특정 값에 매핑될 수 있다. 예를 들어, ASCII 값 64에 대응하는 품질 점수 시퀀스(302)의 '@' 문자는 정수 값 31에 매핑될 수 있다. 유사하게, ASCII 값 65에 대응하는 주어진 품질 점수 시퀀스 내의 'A' 문자는 정수 값 33에 매핑될 수 있는 등이다.

[0086] 미리 결정된 단일 매핑(324)에 도시된 바와 같이, 33의 값에 대응하는 ASCII 문자 '!'는 0과 63 사이인 0의 값에 매핑된다. 34의 값에 대응하는 ASCII 문자 ' " '는 1의 값에 매핑된다. 유사하게, 95의 값에 대응하는 ASCII 문자 '_'는 62의 값에 매핑되고, 96의 값에 대응하는 ASCII 문자 '`'는 63의 값에 매핑된다. 미리 결정된 단일 매핑(324)에 도시되지 않은 다른 매핑들도 단일 매핑 엔진(320)에 의해 생성될 수 있다.

[0087] 다른 구현예들에서, 다른 매핑들이 사용될 수 있다. 예를 들어, 0 내지 63 대신에, 33 내지 96의 ASCII 값들과 같은 품질 점수 시퀀스(302)의 값들이 33 내지 96의 값들에 매핑되는 더 작거나 더 큰 범위들이 사용될 수 있다. 그룹 매핑 엔진(326)에 의해 생성된 매핑들과 같은 다른 매핑들은 33 내지 96의 값들 이외의 값 범위들을 점유할 수 있다. 일부 구현예들에서, 품질 점수 시퀀스(302)의 고유 품질 점수들의 개수는 품질 점수 시퀀스(302)의 품질 점수들을 매핑할 범위를 결정하는 데 사용된다. 예를 들어, 품질 점수 시퀀스(302)가 63개의 고유 품질 점수들을 포함하는 경우, 품질 점수 시퀀스(302)의 품질 점수들을 매핑할 범위는 63개의 값들을 포함할 수 있다. 일부 구현예들에서, 다른 범위들이 사용된다. 예를 들어, 품질 점수 시퀀스(302)가 제1 개수의 고유 품질 점수들을 포함하는 경우, 범위는, 예를 들어, 단일 매핑 엔진(320)에 의해, 고유 품질 점수들의 제1 개수를 2로 나눈 값 또는 고유 품질 점수들의 제1 개수에 기초한 다른 계산을 포함하도록 계산될 수 있다. 일부 구현예들에서, 매핑 범위를 결정하는 데 사용되는 처리 결과에 후속 동작들이 적용된다. 예를 들어, 고유 품질 점수들의 제1 개수가 홀수이고 고유 품질 점수들의 제1 개수에 적용된 제1 동작이 정수 2로 나누는 것인 경우, 제2 동작은, 구현예에 따라, 대응하는 처리 결과를 올림(rounding up) 또는 내림하는 것을 포함할 수 있다.

[0088] 단계 F에서, 그룹 매핑 엔진(326)은 품질 점수 그룹 빈도들(314)을 획득한다. 그룹 매핑 엔진(326)은, 항목(327)에 도시된 바와 같이, 품질 점수 그룹 빈도들(314) 또는 품질 점수 시퀀스(302)와 관련된 다른 데이터를 그룹 매핑 문자 목록(328)과 함께 사용하여, 미리 결정된 그룹 매핑(330)을 생성한다. 도 3의 예에서, 그룹 매핑 문자 목록(328)은 190개의 가장 빈번하게 발생하는 그룹들에 대응하는 정수 값들 64 내지 245를 포함한다. 일부 구현예들에서, 더 많거나 더 적은 그룹들이 그룹 매핑 엔진(326)에 의해 매핑될 수 있다. 예를 들어, 190개의 가장 빈번하게 발생하는 그룹들을 포함하는 미리 결정된 그룹 매핑(330)을 생성하는 대신에, 그룹 매핑 엔진(326)은 200개, 230개, 185개 또는 임의의 다른 개수의 가장 빈번하게 발생하는 그룹들을 포함하는 매핑을 생성할 수 있다. 품질 점수 그룹 빈도들(314)에 대응하는 ASCII 값들은 값들 33 내지 96을 포함한다. 그룹 매핑

엔진(326)은, 품질 점수 그룹 빈도들(314)에 기초하여, 매핑을 위한 그룹들의 일부분을 결정할 수 있다. 예를 들어, 그룹 매핑 엔진(326)은 특정 개수의 가장 빈번하게 발생하는 그룹들(예를 들어, 190개의 가장 빈번하게 발생하는 그룹들)을 결정하고 가장 빈번하게 발생하는 그룹들의 각각의 그룹에 값, 예를 들어, 64 내지 254의 정수 값을 할당할 수 있다.

[0089] 미리 결정된 그룹 매핑(330)에 도시된 바와 같이, 도 3의 예에서, 문자들 'ACD'로 표현된 품질 점수 그룹은 값 64에 매핑된다. 문자들 'FFF'로 표현된 품질 점수 그룹은 값 72에 매핑된다. 문자들 'HIJ'로 표현된 품질 점수 그룹은 값 73에 매핑된다. 미리 결정된 그룹 매핑(330)에 도시되지 않은 다른 매핑들도 그룹 매핑 엔진(326)에 의해 생성될 수 있다.

[0090] 다른 구현예들에서, 다른 매핑들이 사용될 수 있다. 예를 들어, 64 내지 254 대신에, 품질 점수 그룹 빈도들(314) 내의 그룹들이 매핑되는 더 작거나 더 큰 범위들이 사용될 수 있다. 예를 들어, 품질 점수 그룹 빈도들(314) 내의 그룹들은 0 내지 255 사이의 값들에 매핑될 수 있다. 단일 매핑 엔진(320)에 의해 생성된 매핑들과 같은 다른 매핑들이 0 내지 255의 그룹 매핑들에 사용된 값들 이외의 값들을 점유할 수 있다.

[0091] 축소된 시퀀스 생성 엔진(316)은 그룹 매핑 엔진(326) 및 단일 매핑 엔진(320)을 사용하여, 축소된 시퀀스(332)를 생성하기 위해 품질 점수 시퀀스(302)를 처리한다. 품질 점수 시퀀스(302)로부터 도출된 데이터는 품질 점수 시퀀스(332) 자체를 포함할 수 있다. 대안적으로, 품질 점수 시퀀스(302)로부터 도출된 데이터는 빈도 카운터 엔진(310)에 의해 출력된 데이터를 포함할 수 있다. 빈도 카운터 엔진(310)에 의해 출력된 데이터는 품질 점수 그룹 빈도들(314)을 포함할 수 있다. 축소된 시퀀스(332)는 미리 결정된 단일 매핑(324) 및 미리 결정된 그룹 매핑(330)으로부터의 값들의 조합이다.

[0092] 축소된 시퀀스 생성 엔진(316)은 품질 점수 시퀀스(302), 품질 점수 그룹 빈도들, 또는 둘 모두를 처리하고, 축소된 시퀀스(332) 내의 엔트리를 생성하기 위해 단일 매핑 엔진(320) 또는 그룹 매핑 엔진(326)을 사용할지를 결정할 수 있다. 미리 결정된 그룹 매핑(330) 내의 그룹에 속하는 품질 점수 시퀀스(302) 내의 그룹들의 발생들은 미리 결정된 그룹 매핑(330)으로부터의 값으로 대체된다. 예를 들어, 'A' 품질 점수 다음에 'C' 품질 점수가 뒤따르고 'D' 품질 점수가 뒤따르는 발생들은 축소된 시퀀스(332) 내의 값 64로 대체된다.

[0093] 미리 결정된 그룹 매핑(330)의 그룹들 중 하나의 멤버가 아닌 품질 점수들에 대해, 미리 결정된 단일 매핑(324)이 사용된다. 예를 들어, 품질 점수 시퀀스(302)의 첫 번째 값은 ASCII 값 64에 대응하는 문자 '@'이다. 품질 점수 시퀀스(302) 내에서 나타나는 문자 '@'는 미리 결정된 그룹 매핑(330) 내의 그룹의 멤버가 아니다. 이어서, 문자 '@'는 미리 결정된 단일 매핑(324)에 기초하여 매핑된다. 미리 결정된 단일 매핑(324)에 기초한 '@'에 대한 매핑은 값 31이다. 축소된 시퀀스 생성 엔진(316)은 축소된 시퀀스(332) 내에서 '@'의 위치에 대응하는 값 31을 추가한다.

[0094] 단계 G에서, 축소된 시퀀스 생성 엔진(316)은, 위에서 설명되고 항목(331)에 도시된 바와 같이, 미리 결정된 그룹 매핑(330), 미리 결정된 단일 매핑(324), 및 품질 점수 시퀀스(302)로부터 도출된 데이터를 사용하여 축소된 시퀀스(332)를 생성한다. 축소된 시퀀스 생성 엔진(316)은 축소된 시퀀스(332)를 압축 엔진(334)으로 전송한다.

[0095] 단계 H에서, 압축 엔진(334)은 축소된 시퀀스(332)에 대해 미리 결정된 압축 방법에 대응하는 후속 압축을 수행한다. 예를 들어, 일부 경우들에서, 압축 엔진(334)은 축소된 시퀀스(332)를 압축하기 위해 범위 인코더의 PPMD 구현에 기초하여 압축을 수행한다. 일부 구현예들에서, 당업계에 공지된 다른 압축 방법들이 PPMD 구현 대신에 또는 이에 추가하여 사용될 수 있다. 일반적으로, 압축 엔진(334)은 당업계에 공지된 임의의 압축 방법과 같은 임의의 압축 또는 압축 방법들의 조합을 사용할 수 있다. 압축 엔진(334)은 압축 엔진(334)에 의해 사용되는 압축 방법에 대한 입력으로서 축소된 시퀀스(332)에 기초하여 출력(336)을 생성한다. 출력(336)은 품질 점수 시퀀스(302)의 압축된 버전을 표현한다.

[0096] 일부 구현예들에서, 압축 엔진(334)에 공급되는 축소된 시퀀스(332)의 각각의 값은 1바이트 컨텍스트에 따라 압축된다. 예를 들어, 축소된 시퀀스(332)의 주어진 값에 대해, 축소된 시퀀스(332)의 주어진 값을 압축하기 위해 축소된 시퀀스(332) 내의 이전 값이 컨텍스트로서 사용될 수 있다. 1바이트 컨텍스트는 범위 인코더의 PPMD 구현 또는 다른 압축 또는 인코딩 방법에 따라, 축소된 시퀀스(332)를 압축하는 데 사용될 수 있다. 속도, 결과적인 압축 출력, 또는 다른 파라미터들의 균형을 맞추기 위해 더 크거나 더 작은 컨텍스트가 사용될 수 있다. 일부 경우들에서, 출력(336)에 대한 압축비는 품질 점수당 2비트 이상일 수 있다. 예를 들어, 4개 이상의 품질 점수들이 1바이트 또는 8비트의 메모리 공간으로 압축될 수 있다. 일반적으로, 출력(336)은, 도 1의 출력(128)과 유사하게, 구현예에 따라 임의의 추가 프로세스들에서 사용될 수 있거나 저장될 수 있다.

[0097] 설명의 편의를 위해 단계 A 내지 단계 H가 사용된다. 시스템(300)에 의해 수행되는 프로세스는 단계 A 내지 단계 H에 의해 표시된 순서로 발생할 수 있지만, 다른 구현예들에서, 특정 단계들의 순서는 상이할 수 있다. 일부 구현예들에서, 2개 이상의 단계가 동시에 발생할 수 있다.

[0098] 도 4는 제2 데이터 포맷을 갖는 품질 점수들의 시퀀스를 압축하기 위한 프로세스의 예를 예시하는 흐름도이다. 프로세스(400)는 하나 이상의 전자 시스템들, 예를 들어 도 3의 시스템(300)에 의해 수행될 수 있다.

[0099] 프로세스(400)는 유전자 시퀀싱 디바이스에 의해 생성된 유전자 판독 시퀀스를 획득하는 것(402) 및 유전자 판독 시퀀스에 대응하는 복수의 품질 점수들을 획득하는 것(404)을 포함한다. 예를 들어, 도 3에 도시된 바와 같이, 품질 점수 시퀀스(302)는 유전자 데이터 시퀀스에 기초하여 시퀀서에 의해 생성되고 그룹 식별 엔진(304)으로 전송된다. 도 3의 예에서, 품질 점수 시퀀스(302)는 위에서 논의된 바와 같이 염기들과 관련된 품질 점수들을 인코딩하기 위해 Q40 데이터 포맷을 사용하는 시퀀서에 의해 생성된다. 일반적으로, 임계 수 초과의 고유 품질 점수들을 사용하는 임의의 시퀀서는 도 3에 도시된 시스템(300)에 의해 수행되는 프로세스 또는 유사한 프로세스를 사용하여 결과적인 출력 데이터를 압축할 수 있다. 예를 들어, 시퀀서가 유전자 시퀀스에 대응하는 품질 점수들을 인코딩하기 위해 8개 초과의 고유 품질 점수들을 사용하는 경우, 시퀀서는 도 3을 참조하여 논의된 인코딩 및 압축 프로세스들을 사용할 수 있다.

[0100] 프로세스(400)는 복수의 품질 점수들 내의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정하는 것(406)을 포함한다. 예를 들어, 도 3에 도시된 바와 같이, 그룹 식별 엔진(304)은 품질 점수 시퀀스(302)를 획득하고 품질 점수 시퀀스(302) 내의 복수의 품질 점수들에 기초하여 다수의 품질 점수 그룹들을 생성한다. 일부 구현예들에서, 품질 점수들의 그룹들은 품질 점수 시퀀스(302) 내에서 서로 인접한 품질 점수들을 포함한다. 예를 들어, 품질 점수 시퀀스(302)가 품질 점수들 "@ C @ D F F F F..."를 포함하는 경우, 그룹 식별 엔진(304)에 의해 결정된 그룹은 3개의 품질 점수들이 그룹을 생성하는 데 사용되는 구현예들에서 품질 점수들 "@ C @"을 포함할 수 있다.

[0101] 일부 구현예들에서, 그룹 식별 엔진(304)은 하나 이상의 그룹들을 빈도 카운터 엔진(310)으로 전송하고 빈도 카운터 엔진(310)은 각각의 품질 점수 그룹의 발생 빈도를 결정한다. 예를 들어, 빈도 카운터 엔진(310)은, 그룹 식별 엔진(304)에 의해 결정된 각각의 그룹에 대해, 그룹의 얼마나 많은 발생들이 품질 점수 시퀀스(302) 내에 존재할 수 있는지를 결정할 수 있다. 도 3의 예에서, 품질 점수 시퀀스(302)의 품질 점수들은 3개의 그룹들로 그룹화된다. 그러나, 다른 구현예들에서는, 다른 개수의 품질 점수들이 사용될 수 있다. 예를 들어, 그룹 식별 엔진(304)은 품질 점수 그룹을 생성하기 위해 4개의 품질 점수들, 2개의 품질 점수들, 또는 임의의 다른 수의 품질 점수들의 그룹들을 결정할 수 있다.

[0102] 프로세스(400)는, 복수의 품질 점수들 내의 각각의 품질 점수에 대해, 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버인지 여부를 결정하는 것(408)을 포함한다. 예를 들어, 도 3에 도시된 바와 같이, 축소된 시퀀스 생성 엔진(316), 단일 매핑 엔진(320) 및 그룹 매핑 엔진(326)은 미리 결정된 단일 매핑(324) 및 미리 결정된 그룹 매핑(330)에 기초하여 축소된 시퀀스(332)를 생성한다. 축소된 시퀀스(332) 내의 각각의 품질 점수에 대해, 축소된 시퀀스 생성 엔진(316)은 품질 점수가 미리 결정된 그룹 매핑(330)에 포함된 그룹의 멤버인지 여부를 결정한다. 품질 점수가 미리 결정된 그룹 매핑(330)에 포함된 그룹의 멤버인 경우, 품질 점수는, 그룹의 다른 품질 점수들과 함께, 축소된 시퀀스(332) 내의 단일 엔트리로서 인코딩된다.

[0103] 일부 구현예들에서, 미리 결정된 그룹 매핑(330)의 각각의 그룹은 미리 결정된 임계치를 충족한다. 예를 들어, 미리 결정된 그룹 매핑(330)의 각각의 그룹은 그룹 매핑에 포함될 미리 결정된 횟수만큼 품질 점수 시퀀스(302) 내에서 발생할 수 있다. 미리 결정된 그룹 매핑(330)의 각각의 그룹은 품질 점수 시퀀스(302) 내에서 다른 그룹들보다 더 빈번하게 발생하는 미리 결정된 수의 그룹들에 속할 수 있다. 도 3의 예에서, 미리 결정된 그룹 매핑(330)은 190개의 가장 빈번하게 발생하는 품질 점수 그룹들을 포함한다. 이어서, 임계치는, 구현예에 따라 다른 것들 중에서도, 191번째 가장 빈번하게 발생하는 품질 점수 그룹, 190번째 가장 빈번하게 발생하는 품질 점수 그룹에 대응하는 발생 횟수로서 정의될 수 있다. 이어서, 미리 결정된 그룹 매핑(330)의 190개의 가장 빈번하게 발생하는 품질 점수 그룹들은 모두 이 임계치를 충족한다.

[0104] 일부 구현예들에서, 상이한 양의 품질 점수 그룹들이 미리 결정된 그룹 매핑(330)을 생성하는 데 사용될 수 있다. 예를 들어, 미리 결정된 그룹 매핑(330)의 그룹들의 개수는 시퀀서에 의해 사용되는 품질 점수들의 개수에 기초하여 결정될 수 있다. 이어서, 매핑은 최적화 프로세스를 충족하도록 생성될 수 있다. 예를 들어, 최적화 프로세스는 단일 값으로서 인코딩되는 그룹들의 개수를 최대화하는 것, 및 미리 결정된 그룹 매핑(330)의 품질

점수 그룹에 속하지 않고 단일 값으로서 인코딩되는 품질 점수들의 개수를 최소화하는 것을 포함할 수 있다. 최적화할 파라미터들은 도 3의 미리 결정된 그룹 매핑(330)과 같은 그룹 매핑에 포함시킬 그룹들의 개수뿐만 아니라, 미리 결정된 그룹 매핑(330) 내에서 하나 이상의 그룹들을 생성하는 데 사용되는 품질 점수들의 개수를 포함할 수 있다. 도 3의 미리 결정된 단일 매핑(324)과 같은 단일 매핑 내의 품질 점수들의 개수는 유전자 시퀀스 판독에 대응하는 품질 점수들을 표현하기 위해 시퀀서에 의해 사용되는 고유 품질 점수들의 개수에 대응할 수 있다.

[0105] 일부 구현예들에서, 품질 점수 그룹들은 품질 점수 시퀀스 내에서 함께 나타나는 인접한 품질 점수들에 기초한다. 예를 들어, "@ C @ D F F F F…"로 표현되는 품질 점수 시퀀스(302)는 인접한 품질 점수들의 그룹 "FFF"를 포함한다. 인접한 품질 점수들의 그룹 "FFF"는 미리 결정된 그룹 매핑(330)에 포함된다. 미리 결정된 그룹 매핑(330)에 따르면, 인접한 품질 점수들의 그룹 "FFF"는 축소된 시퀀스 생성 엔진(316)에 의해 단일 값 72로서 인코딩되어야 한다. 물론, 다른 구현예들에서, 매핑 또는 인코딩 목적을 위해 다른 값들이 사용될 수 있다.

[0106] 일부 구현예들에서, 품질 점수 그룹들의 멤버들이 아닌 점수들은 축소된 시퀀스 내에서 단일 엔트리로서 매핑된다. 예를 들어, 품질 점수들의 그룹을 생성하기 위해 3개의 품질 점수들을 사용하는 구현예에서, "@ C @ D F F F F…"로 표현되는 품질 점수 시퀀스(302)의 첫 번째 값 "@"는 특정 품질 점수 그룹의 멤버가 아니다. 이와 같이, 첫 번째 값 "@"는 미리 결정된 단일 매핑(324)에 따라 단일 값 31로서 축소된 시퀀스(332)에서 인코딩된다. 도 3의 예에서, 미리 결정된 단일 매핑(324)은, 품질 점수 시퀀스(302) 내의 품질 점수들을 표현하는 데 사용되는 값들과, 동일한 길이의 새로운 범위의, 일대일 매칭이므로, 미리 결정된 그룹 매핑(330) 내의 그룹의 일부가 아닌 임의의 품질 점수가, 미리 걸정된 단일 매핑(324)을 생성하는 데 사용되는 시프팅 값을 원래 값에 더하거나 뺀 값으로서, 축소된 시퀀스(332)에서 인코딩되도록 한다. 예를 들어, 도 3에서, 품질 점수 그룹의 멤버가 아닌 품질 점수들에 대응하는 단일 값들은 33의 값만큼 시프팅된다. 시프팅은 그룹 매핑들 또는 단일 매핑들에 할애된 값들의 연속적인 범위를 생성하는 데 사용할 수 있다.

[0107] 프로세스(400)는 품질 점수가 특정 품질 점수 그룹의 멤버라는 결정에 기초하여 특정 품질 점수 그룹의 표현으로서 축소된 시퀀스에 포함시키기 위한 단일 엔트리를 생성하는 것(410)을 포함한다. 예를 들어, 축소된 시퀀스 생성 엔진(316)은 "@ C @ D F F F F…"로 표현되는 품질 점수 시퀀스(302)가 품질 점수들의 그룹 "FFF"를 포함한다고 결정할 수 있다. 품질 점수들의 그룹 "FFF"는 미리 결정된 그룹 매핑(330)에 포함된다. 미리 결정된 그룹 매핑(330)에 따르면, 인접한 품질 점수들의 그룹 "FFF"는 축소된 시퀀스 생성 엔진(316)에 의해 단일 값 72로서 인코딩되어야 한다. 그러나, 본 개시내용은 품질 점수들의 그룹들을 단일 엔트리 또는 값으로 대체하는 것으로 제한되지 않는다. 다른 구현예들에서, 단일 품질 점수는 미리 결정된 단일 매핑을 사용하여 단일 엔트리 또는 값으로 대체될 수 있다. 이러한 단일 품질 점수 대체들은, 단일 엔트리 또는 값이 단일 품질 점수보다 더 작은 비트 크기를 갖는 단일 품질 점수를 표현하도록 선택될 수 있기 때문에, 유리할 수 있다.

[0108] 프로세스(400)는 생성된 엔트리들 각각을 집계함으로써 축소된 시퀀스를 생성하는 것(412)을 포함한다. 예를 들어, 축소된 시퀀스 생성 엔진(316)은 "@ C @ D F F F F…"로 표현된 품질 점수 시퀀스(302)의 품질 점수 "@"에 대응하는 31의 제1 인코딩된 값, 품질 점수 "C"에 대응하는 34의 제2 인코딩된 값, 품질 점수 "@"에 대응하는 31의 제3 인코딩된 값, 품질 점수 "D"에 대응하는 35의 추가 인코딩된 값, 및 품질 점수 그룹 "FFF"에 대응하는 72의 제5 인코딩된 값을 집계할 수 있다. 축소된 시퀀스 생성 엔진(316)은 품질 점수 시퀀스(302)의 후속 값들에 기초하여 후속 인코딩된 값들을 생성할 수 있다. 축소된 시퀀스 생성 엔진(316)은 품질 점수 시퀀스(302)의 모든 값들이 축소된 시퀀스(332) 내에서 표현될 때까지 계속될 수 있다.

[0109] 일부 구현예들에서, 축소된 시퀀스(332)에 기초하여 추가 압축 단계들이 수행된다. 예를 들어, 도 3에 도시된 바와 같이, 축소된 시퀀스(332)는 압축 엔진(334)으로 전송될 수 있다. 이어서, 압축 엔진(334)은 출력(336)을 생성하기 위해 축소된 시퀀스(332)에 대해 하나 이상의 압축 프로세스들을 수행할 수 있다. 도 1에 도시된 프로세스와 유사하게, 압축 엔진(334)에 의해 수행되는 후속 압축 단계들 이전에 생성되는 축소된 시퀀스(332)로 표현되는 첫 번째 인코딩은 출력(336)의 결과적인 압축을 개선하는 데 유리할 수 있다. 축소된 시퀀스(332)의 형태는 압축 엔진(334)이 원래의 품질 점수 시퀀스(302)보다 더 빠르게 또는 효과적으로 데이터를 압축할 수 있도록 할 수 있다. 예를 들어, 축소된 시퀀스(332)는 품질 점수 시퀀스(302)의 압축된 버전일 수 있고, 압축 엔진(334)에 의해 수행되는 압축의 지속시간 또는 품질은 입력의 크기에 의존할 수 있다. 이러한 방식으로, 축소된 시퀀스(332)를 생성하기 위한 초기 인코딩 단계들은 압축하는 데 걸리는 시간을 감소시키고 압축 엔진(334)에 의해 달성되는 압축의 품질을 증가시킬 수 있다.

[0110] 도 5는 제1 데이터 포맷을 갖는 품질 점수들의 시퀀스를 압축해제하기 위한 프로세스(500)의 예를 예시하는 흐

름도이다. 프로세스(500)는 하나 이상의 전자 시스템들, 예를 들어 도 1의 시스템(100)에 의해 수행될 수 있다.

[0111] 프로세스(500)는 기수-x 수를 사용하여 복수의 품질 점수들 각각을 인코딩함으로써 생성된 제1 인코딩된 데이터 세트를 획득하는 것을 포함하며, 여기서 x는 유전자 시퀀싱 디바이스에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수이다(502). 예를 들어, 디코딩 엔진은 이진 출력(114) 또는 이진 출력(126)을 획득할 수 있다.

[0112] 프로세스(500)는 기수-x 수를 사용하여 제1 디코딩된 데이터 세트를 생성하는 것(504)을 포함한다. 예를 들어, 항목들(108, 120)에 도시된 프로세스들과 유사하게, 제1 인코딩된 데이터 세트는 인코딩에 사용된 기수-x 수에 기초하여 디코딩될 수 있으며, 여기서 x는 복수의 품질 점수들에 존재하는 고유 품질 점수들의 개수에 대응하는 정수이다. 일부 구현예들에서, 제1 디코딩된 데이터의 이진 표현의 정수 값은 기수-x 수를 생성하기 위해 기수-x 수의 x로 반복적으로 나누어질 수 있다.

[0113] 프로세스(500)는 하나 이상의 다른 디코딩된 데이터 세트들 내에서 제1 디코딩된 데이터 세트를 순서화하는 것(506)을 포함한다. 일부 구현예들에서, 순서화 엔진은 디코딩된 데이터 세트 및 하나 이상의 다른 디코딩된 데이터 세트들을 획득하고 제1 인코딩된 데이터 세트에 기초하여 제1 디코딩된 데이터를 순서화할 수 있다. 예를 들어, 순서화 엔진은, 제1 디코딩된 데이터 세트를 생성하기 위해 이어서 디코딩되는, 제1 인코딩된 데이터 세트의 일부분을 결정할 수 있다. 순서화 엔진은 또한 하나 이상의 다른 디코딩된 데이터 세트들을 생성하기 위해 디코딩되는 제1 인코딩된 데이터 세트의 부분들을 결정할 수 있다. 제1 인코딩된 데이터 세트 내의 제1 디코딩된 데이터 세트 및 하나 이상의 다른 디코딩된 데이터 세트들의 원래 위치에 기초하여, 순서화 엔진은 하나 이상의 다른 디코딩된 데이터 세트들 내에서 제1 디코딩된 데이터 세트를 순서화할 수 있다.

[0114] 프로세스(500)는 제1 디코딩된 데이터 세트 및 하나 이상의 다른 디코딩된 데이터 세트들에 기초하여 집계(aggregate) 디코딩된 데이터 세트를 생성하는 것(508)을 포함한다. 예를 들어, 하나 이상의 다른 디코딩된 데이터 세트들 내의 제1 디코딩된 데이터 세트의 순서화에 기초하여, 집계 디코딩된 데이터 세트 엔진은 제1 디코딩된 데이터 세트 및 하나 이상의 다른 디코딩된 데이터 세트들을 포함하는 집계 디코딩된 데이터 세트를 생성할 수 있다. 집계 디코딩된 데이터 세트는 제1 인코딩된 데이터 세트를 생성하는 데 사용된 데이터와 유사한 데이터를 포함할 수 있다. 집계 디코딩된 데이터 세트는 구현예에 따라 다른 프로세스들 또는 동작들에서 사용될 수 있다.

[0115] 도 6은 제2 데이터 포맷을 갖는 품질 점수들의 시퀀스를 압축해제하기 위한 프로세스(600)의 예를 예시하는 흐름도이다. 프로세스(600)는 하나 이상의 전자 시스템들, 예를 들어 도 3의 시스템(300)에 의해 수행될 수 있다.

[0116] 프로세스(600)는 단일 매핑 데이터베이스 및 그룹 매핑 데이터베이스에 기초하여 생성된 축소된 시퀀스를 획득하는 것(602)을 포함한다. 예를 들어, 디코딩 엔진은 시스템(300) 또는 다른 시스템으로부터 축소된 시퀀스(332) 또는 다른 관련 데이터를 획득할 수 있다. 이어서, 축소된 시퀀스 생성 엔진(316)에 대응하여 도시된 프로세스는 품질 점수 시퀀스(302)를 생성하기 위해 역으로 수행될 수 있다.

[0117] 프로세스(600)는 축소된 시퀀스 및 단일 매핑 데이터베이스 및 그룹 매핑 데이터베이스에 기초하여 제1 디코딩된 데이터 세트를 생성하는 것(604)을 포함한다. 위에서 논의된 바와 같이, 디코딩 엔진은 품질 점수 시퀀스(302)를 생성하기 위해 축소된 시퀀스 생성 엔진(316)과 유사하지만 역순으로 동작들을 수행할 수 있다. 예를 들어, 디코딩 엔진은 축소된 시퀀스(332)를 획득하고 미리 결정된 그룹 매핑(330) 및 미리 결정된 단일 매핑(324)을 사용하여 축소된 시퀀스(332)를 디코딩하고 품질 점수 시퀀스(302)의 적어도 일부분에 대응하는 제1 디코딩된 데이터 세트를 생성할 수 있다.

[0118] 프로세스(600)는 하나 이상의 다른 디코딩된 데이터 세트들 내에서 제1 디코딩된 데이터 세트를 순서화하는 것(606)을 포함한다. 예를 들어, 디코딩 엔진은 축소된 시퀀스(332)의 제1 부분을 디코딩하고 축소된 시퀀스(332)의 제1 부분에 대응하는 제1 순서를 결정할 수 있다. 일부 구현예들에서, 축소된 시퀀스(332)의 순서는 제1 디코딩된 데이터 세트 및 하나 이상의 다른 디코딩된 데이터 세트들의 순서를 결정하기 위해 사용될 수 있다. 예를 들어, 제1 디코딩된 데이터 세트는 축소된 시퀀스(332)의 제1 부분에 대응할 수 있다. 결과적으로, 제1 디코딩된 데이터 세트는 최종 집계된 디코딩된 데이터 세트의 시작 부분에서 순서화될 수 있다. 이어서, 후속 하나 이상의 다른 디코딩된 데이터 세트들은 하나 이상의 다른 디코딩된 데이터 세트들을 생성하는 데 사용된 축소된 시퀀스(332)에 대응하는 데이터의 순서에 기초하여 순서화될 수 있다.

[0119] 프로세스(600)는 제1 디코딩된 데이터 세트 및 하나 이상의 다른 디코딩된 데이터 세트들에 기초하여 집계 디코딩된 데이터 세트를 생성하는 것(608)을 포함한다. 예를 들어, 집계 디코딩된 데이터 세트 엔진은 위에서 논의

된 바와 같이 하나 이상의 다른 디코딩된 데이터 세트 내에서 제1 디코딩된 데이터 세트를 순서화하는 것에 기초하여 집계 디코딩된 데이터 세트를 생성하기 위해 사용될 수 있다. 집계 디코딩된 데이터 세트는 축소된 시퀀스(332)와 같은 축소된 시퀀스를 생성하는 데 사용된 데이터와 유사한 데이터를 포함할 수 있다. 집계 디코딩된 데이터 세트는 구현예에 따라 다른 프로세스들 또는 동작들에서 사용될 수 있다.

[0120] 도 7은 품질 점수들의 압축 방법을 결정하기 위한 프로세스(700)의 예를 예시하는 흐름도이다. 프로세스(700)는 하나 이상의 전자 시스템들, 예를 들어, 도 3의 시스템(300) 또는 도 1의 시스템(100)에 의해 수행될 수 있다.

[0121] 프로세스(700)는 유전자 시퀀서로부터 유전자 데이터를 획득하는 것(702)을 포함한다. 다양한 형태의 유전자 시퀀서들이 당업자에게 알려져 있을 것이다. 예를 들어, 결정 엔진은 주어진 유전자 시퀀서에 의해 생성된 하나 이상의 품질 점수들에 대응하는 유전자 데이터를 획득할 수 있다.

[0122] 프로세스(700)는 유전자 데이터 내의 고유 품질 점수들의 개수를 결정하는 것(704)을 포함한다. 예를 들어, 주어진 시퀀서의 유형, 모델 또는 특정 소프트웨어에 따라, 염기 호출들에 대한 품질 점수들은 하나 이상의 심볼들 또는 값들에 의해 표현될 수 있다. 이어서, 하나 이상의 품질 점수들을 표현하는 데 사용되는 고유 심볼들 또는 값들의 개수는, 유전자 데이터의 품질 점수들에 사용할 압축 방법을 결정하는 데 사용될 수 있다.

[0123] 프로세스(700)는 유전자 시퀀서로부터 획득된 유전자 데이터 내에 8개 이하의 고유 품질 점수들이 존재한다는 결정에 대응하는 제1 결정 경로 및 유전자 시퀀서로부터 획득된 유전자 데이터 내에 8개 초과의 고유 품질 점수들이 존재한다는 결정에 대응하는 제2 결정 경로를 포함한다. 8개 이하의 고유 품질 점수들이 유전자 데이터 내에 존재하는 경우, 도 2에서 설명된 프로세스(200)에 대응하는 프로세스가 유전자 데이터의 품질 점수들을 압축하기 위해 수행될 수 있다. 8개 초과의 고유 품질 점수들이 유전자 데이터 내에 존재하는 경우, 도 4에서 설명된 프로세스(400)에 대응하는 프로세스가 유전자 데이터의 품질 점수들을 압축하기 위해 수행될 수 있다.

[0124] 일부 구현예들에서, 사용할 압축 방법을 결정하기 위해 다른 임계치들이 사용된다. 예를 들어, 8개의 고유 품질 점수들 대신에, 시스템은 도 4에 설명된 프로세스(400)에 대응하는 프로세스를 추구하기 위해 7, 9, 또는 10개의 고유 품질 점수들이 필요하다고 결정할 수 있다. 일반적으로, 시스템은 주어진 품질 점수들의 세트와 같은 주어진 데이터 세트를 압축하기 위해 복수의 상이한 압축 방법들 중 어느 것을 사용할지를 결정하기 위해 임의의 적합한 임계치들 또는 결정 모델링을 사용할 수 있다.

[0125] 본 개시내용에서 지칭되는 게놈 데이터, 예를 들어, 입력 데이터(102) 등은, 제한으로서가 아니라, 예를 들어, 뉴클레오티드 시퀀스, 데옥시리보핵산(DNA) 시퀀스, 리보핵산(RNA), 및 아미노산 시퀀스를 포함할 수 있다. 본 명세서의 설명이 뉴클레오티드 시퀀스 형태의 게놈 정보와 관련하여 상당히 상세하게 설명되어 있지만, 본 명세서의 순서화된 데이터 시퀀스는, 수 개의 변형들이 있기는 하지만, 다른 게놈 데이터에 대해서도 역시 구현될 수 있음이 이해될 것이며, 이는 당업자에 의해 이해되는 바와 같다.

[0126] 도 8은 제1 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩, 예를 들어, 압축하기 위한 프로세스의 실험 결과들의 그래픽 묘사이다. 도 8은 데이터 세트 "SRR6882909_1.fastq"의 압축으로부터의 결과들을 도시한다. 데이터 세트 "SRR6882909_1.fastq"는 위에서 설명된 Q4 포맷을 사용하여 포맷화된다. 데이터 세트 "SRR6882909_1.fastq"의 품질 점수들은 4개의 고유 품질 점수들을 포함한다. 차트(802)는 데이터 세트 "SRR6882909_1.fastq"에 대응하는 원시 데이터뿐만 아니라, gzipped 레벨 9 압축, zstd 레벨 11 압축, 및 Lena Q4 압축을 포함하는 다양한 기법들을 사용한 압축 후의 데이터 세트 "SRR6882909_1.fastq"의 압축된 버전들의 상대적 크기들을 도시한다. Lena Q4 압축은 도 1에 도시된 프로세스(100) 및 본 명세서에 설명된 방법들에 대응한다.

[0127] 차트(802)는 데이터 세트 "SRR6882909_1.fastq"의 원시 데이터가 7402335856바이트이고, gzipped 레벨 9 압축을 사용한 데이터 세트 "SRR6882909_1.fastq"의 압축된 버전은 417512395바이트이고, zstd 레벨 11 압축을 사용한 데이터 세트 "SRR6882909_1.fastq"의 압축된 버전은 452733689바이트이고, Lena Q4 압축을 사용한 데이터 세트 "SRR6882909_1.fastq"의 압축된 버전은 259865991바이트임을 도시한다. 데이터 세트 "SRR6882909_1.fastq"의 Lena Q4 압축에 의해 획득된 압축은 다른 대안적인 압축 방법들에 의해 달성된 압축보다 더 크다. 범례(806)는 어느 막대가 어느 압축 방법에 대응하고 어느 막대가 데이터 세트 "SRR6882909_1.fastq"의 원시 데이터에 대응하는지를 도시한다.

[0128] 차트(804)는 데이터 세트 "SRR6882909_1.fastq"에 대해 사용된 압축 방법들 각각에 대한 압축 시간들을 도시한다. 차트(804)는 gzipped 레벨 9 압축을 사용한 데이터 세트 "SRR6882909_1.fastq"의 압축된 버전이 생성하는 데 1980초가 걸리고, zstd 레벨 11 압축을 사용한 데이터 세트 "SRR6882909_1.fastq"의 압축된 버전이 생성하는

데 108초가 걸리고, Lena Q4 압축을 사용한 데이터 세트 "SRR6882909_1.fastq"의 압축된 버전이 생성하는 데 36초가 걸린다는 것을 도시한다. Lena Q4 압축을 사용하여 데이터 세트 "SRR6882909_1.fastq"를 압축하는 시간은 다른 대안적인 압축 방법들에 의해 달성된 압축 시간보다 더 작다.

[0129] 도 9는 제2 데이터 포맷을 갖는 품질 점수들의 시퀀스를 인코딩, 예를 들어, 압축하기 위한 프로세스의 실험 결과들의 그래픽 묘사이다. 도 9는 데이터 세트 "ERR1744700_1.fastq"의 압축으로부터의 결과들을 도시한다. 데이터 세트 "ERR1744700_1.fastq"는 위에서 설명된 Q40 포맷을 사용하여 포맷화된다. 데이터 세트 "ERR1744700_1.fastq"의 품질 점수들은 4개 초과의 고유 품질 점수들을 포함한다. 차트(902)는 데이터 세트 "ERR1744700_1.fastq"에 대응하는 원시 데이터뿐만 아니라, gzipped 레벨 9 압축, zstd 레벨 11 압축, 및 Lena Q40 압축을 포함하는 다양한 기법들을 사용한 압축 후의 데이터 세트 "ERR1744700_1.fastq"의 압축된 버전들의 상대적 크기들을 도시한다. Lena Q40 압축은 도 3에 도시된 프로세스(300)에 대응한다.

[0130] 차트(902)는 데이터 세트 "ERR1744700_1.fastq"의 원시 데이터가 5033592178바이트이고, gzipped 레벨 9 압축을 사용한 데이터 세트 "ERR1744700_1.fastq"의 압축된 버전은 1289564690바이트이고, zstd 레벨 11 압축을 사용한 데이터 세트 "ERR1744700_1.fastq"의 압축된 버전은 1290828665바이트이고, Lena Q40 압축을 사용한 데이터 세트 "ERR1744700_1.fastq"의 압축된 버전은 1228518456바이트임을 도시한다. 데이터 세트 "ERR1744700_1.fastq"의 Lena Q40 압축에 의해 획득된 압축은 다른 대안적인 압축 방법들에 의해 달성된 압축보다 더 크다. 범례(906)는 어느 막대가 어느 압축 방법에 대응하고 어느 막대가 데이터 세트 "ERR1744700_1.fastq"의 원시 데이터에 대응하는지를 도시한다.

[0131] 차트(904)는 데이터 세트 "ERR1744700_1.fastq"에 대해 사용된 압축 방법들 각각에 대한 압축 시간들을 도시한다. 차트(904)는 gzipped 레벨 9 압축을 사용한 데이터 세트 "ERR1744700_1.fastq"의 압축된 버전이 생성하는 데 3762초가 걸리고, zstd 레벨 11 압축을 사용한 데이터 세트 "ERR1744700_1.fastq"의 압축된 버전이 생성하는 데 306초가 걸리고, Lena Q40 압축을 사용한 데이터 세트 "ERR1744700_1.fastq"의 압축된 버전이 생성하는 데 101초가 걸린다는 것을 도시한다. Lena Q40 압축을 사용하여 데이터 세트 "ERR1744700_1.fastq"를 압축하는 시간은 다른 대안적인 압축 방법들에 의해 달성된 압축 시간보다 더 작다.

[0132] 도 10은 다변량 서수 데이터에 기초한 조인트 모델을 사용하여 의료 분석을 생성하기 위한 시스템을 구현하는 데 사용될 수 있는 컴퓨터 시스템(1000) 컴포넌트들의 도면이다.

[0133] 컴퓨팅 디바이스(1000)는 랩톱, 데스크톱, 워크스테이션, 개인 휴대 정보 단말기, 서버, 블레이드 서버, 메인프레임, 및 다른 적합한 컴퓨터와 같은 다양한 형태들의 디지털 컴퓨터들을 표현하도록 의도된다. 컴퓨팅 디바이스(1050)는 개인 휴대 정보 단말기, 셀룰러 전화, 스마트폰, 및 다른 유사한 컴퓨팅 디바이스들과 같은 다양한 형태들의 모바일 디바이스들을 표현하도록 의도된다. 추가적으로, 컴퓨팅 디바이스(1000 또는 1050)는 범용 직렬 버스(Universal Serial Bus, USB) 플래시 드라이브들을 포함할 수 있다. USB 플래시 드라이브들은 운영 체제들 및 다른 애플리케이션들을 저장할 수 있다. USB 플래시 드라이브들은 다른 컴퓨팅 디바이스의 USB 포트 내로 삽입될 수 있는 무선 송신기 또는 USB 커넥터와 같은 입력/출력 컴포넌트들을 포함할 수 있다. 본 명세서에 도시된 컴포넌트들, 그들의 접속들 및 관계들, 및 그들의 기능들은 단지 예들일 뿐이며, 본 명세서에서 설명되고/되거나 청구되는 본 발명들의 구현예들을 제한하고자 하는 것은 아니다.

[0134] 컴퓨팅 디바이스(1000)는 프로세서(1002), 메모리(1004), 저장 디바이스(1008), 메모리(1004) 및 고속 확장 포트들(1010)에 접속되는 고속 인터페이스(1008), 및 저속 버스(1014) 및 저장 디바이스(1008)에 접속되는 저속 인터페이스(1012)를 포함한다. 컴포넌트들(1002, 1004, 1008, 1008, 1010, 1012) 각각은 다양한 버스들을 사용하여 상호접속되고, 공통 마더보드 상에 또는 적합한 다른 방식으로 장착될 수 있다. 프로세서(1002)는, 고속 인터페이스(1008)에 결합된 디스플레이(1016)와 같은 외부 입력/출력 디바이스 상에 GUI에 대한 그래픽 정보를 디스플레이하기 위해 메모리(1004) 내에 또는 저장 디바이스(1008) 상에 저장된 명령어들을 포함하는, 컴퓨팅 디바이스(1000) 내의 실행을 위한 명령어들을 처리할 수 있다. 다른 구현예들에서, 다수의 프로세서들 및/또는 다수의 버스들이 다수의 메모리들 및 메모리의 유형들과 함께 적절하게 사용될 수 있다. 또한, 다수의 컴퓨팅 디바이스들(1000)이 접속될 수 있는데, 이때 각각의 디바이스는, 예컨대 서버 뱅크, 블레이드 서버들의 그룹, 또는 멀티-프로세서 시스템으로서, 필요한 동작들의 부분들을 제공한다.

[0135] 메모리(1004)는 컴퓨팅 디바이스(1000) 내에 정보를 저장한다. 하나의 구현예에서, 메모리(1004)는 휘발성 메모리 유닛 또는 유닛들이다. 다른 구현예에서, 메모리(1004)는 비휘발성 메모리 유닛 또는 유닛들이다. 메모리(1004)는 또한, 자기 또는 광 디스크와 같은 다른 형태의 컴퓨터 판독가능 매체일 수 있다.

[0136] 저장 디바이스(1008)는 컴퓨팅 디바이스(1000)를 위한 대용량 저장소를 제공할 수 있다. 하나의 구현예에서, 저장 디바이스(1008)는 컴퓨터 판독가능 매체, 예컨대 플로피 디스크 디바이스, 하드 디스크 디바이스, 광 디스크 디바이스, 또는 테이프 디바이스, 플래시 메모리 또는 다른 유사한 솔리드 스테이트 메모리 디바이스, 또는 저장 영역 네트워크 내의 디바이스들 또는 다른 구성들을 포함하는 디바이스들의 어레이일 수 있거나 이들을 포함할 수 있다. 컴퓨터 프로그램 제품이 정보 캐리어 내에 유형적으로 구현될 수 있다. 컴퓨터 프로그램 제품은 또한, 실행될 때 전술된 것들과 같은 하나 이상의 방법들을 수행하는 명령어들을 포함할 수 있다. 정보 캐리어는 메모리(1004), 저장 디바이스(1008), 또는 프로세서(1002) 상의 메모리와 같은 컴퓨터 또는 기계 판독가능 매체이다.

[0137] 고속 제어기(1008)는 컴퓨팅 디바이스(1000)에 대한 대역폭 집약적 동작들을 관리하는 한편, 저속 제어기(1012)는 더 낮은 대역폭 집약적 동작들을 관리한다. 기능들의 그러한 할당은 단지 일례이다. 하나의 구현예에서, 고속 제어기(1008)는, 예컨대 그래픽 프로세서 또는 가속기를 통해 메모리(1004), 디스플레이(1016)에, 그리고 다양한 확장 카드들(도시되지 않음)을 수용할 수 있는 고속 확장 포트들(1010)에 결합된다. 그 구현예에서, 저속 제어기(1012)는 저장 디바이스(1008) 및 저속 확장 포트(1014)에 결합된다. 다양한 통신 포트들, 예컨대 USB, 블루투스, 이더넷, 무선 이더넷을 포함할 수 있는 저속 확장 포트는 하나 이상의 입력/출력 디바이스들, 예컨대 키보드, 포인팅 디바이스, 마이크로폰/스피커 쌍, 스캐너, 또는 스위치나 라우터와 같은 네트워킹 디바이스에, 예를 들어 네트워크 어댑터를 통해 결합될 수 있다. 컴퓨팅 디바이스(1000)는, 도면에 도시된 바와 같이, 다수의 상이한 형태들로 구현될 수 있다. 예를 들어, 그것은 표준 서버(1020)로서, 또는 그러한 서버들의 그룹으로 다수회 구현될 수 있다. 그것은 또한 랙 서버 시스템(1024)의 일부로서 구현될 수 있다. 더욱이, 그것은 랩톱 컴퓨터(1022)와 같은 개인용 컴퓨터에서 구현될 수 있다. 대안적으로, 컴퓨팅 디바이스(1000)로부터의 컴포넌트들은 디바이스(1050)와 같은 모바일 디바이스(도시되지 않음) 내의 다른 컴포넌트들과 조합될 수 있다. 그러한 디바이스들 각각은 컴퓨팅 디바이스(1000, 1050) 중 하나 이상을 포함할 수 있고, 전체 시스템은 서로 통신하는 다수의 컴퓨팅 디바이스들(1000, 1050)로 구성될 수 있다.

[0138] 컴퓨팅 디바이스(1000)는, 도면에 도시된 바와 같이, 다수의 상이한 형태들로 구현될 수 있다. 예를 들어, 그것은 표준 서버(1020)로서, 또는 그러한 서버들의 그룹으로 다수회 구현될 수 있다. 그것은 또한 랙 서버 시스템(1024)의 일부로서 구현될 수 있다. 더욱이, 그것은 랩톱 컴퓨터(1022)와 같은 개인용 컴퓨터에서 구현될 수 있다. 대안적으로, 컴퓨팅 디바이스(1000)로부터의 컴포넌트들은 디바이스(1050)와 같은 모바일 디바이스(도시되지 않음) 내의 다른 컴포넌트들과 조합될 수 있다. 그러한 디바이스들 각각은 컴퓨팅 디바이스(1000, 1050) 중 하나 이상을 포함할 수 있고, 전체 시스템은 서로 통신하는 다수의 컴퓨팅 디바이스들(1000, 1050)로 구성될 수 있다.

[0139] 컴퓨팅 디바이스(1050)는, 다른 컴포넌트들 중에서도, 프로세서(1052), 메모리(1064), 및 입력/출력 디바이스, 예컨대 디스플레이(1054), 통신 인터페이스(1066), 및 송수신기(1068)를 포함한다. 디바이스(1050)에는 또한, 추가적인 저장소를 제공하기 위해, 마이크로 드라이브 또는 다른 디바이스와 같은 저장 디바이스가 제공될 수 있다. 컴포넌트들(1050, 1052, 1064, 1054, 1066, 1068) 각각은 다양한 버스들을 사용하여 상호접속되고, 몇몇 컴포넌트들은 공통 마더보드 상에 또는 적합한 다른 방식으로 장착될 수 있다.

[0140] 프로세서(1052)는 메모리(1064)에 저장된 명령어들을 포함하는, 컴퓨팅 디바이스(1050) 내의 명령어들을 실행시킬 수 있다. 프로세서는 별개의 그리고 다수의 아날로그 및 디지털 프로세서들을 포함하는 칩들의 칩셋으로서 구현될 수 있다. 더욱이, 프로세서는 다수의 아키텍처들 중 임의의 것을 사용하여 구현될 수 있다. 예를 들어, 프로세서(1010)는 CISC(Complex Instruction Set Computer) 프로세서, RISC(Reduced Instruction Set Computer) 프로세서, 또는 MISC(Minimal Instruction Set Computer) 프로세서일 수 있다. 프로세서는, 예를 들어, 디바이스(1050)의 다른 컴포넌트들의 조정, 예컨대 사용자 인터페이스들의 제어, 디바이스(1050)에 의해 실행되는 애플리케이션들, 및 디바이스(1050)에 의한 무선 통신을 제공할 수 있다.

[0141] 프로세서(1052)는 디스플레이(1054)에 결합된 디스플레이 인터페이스(1056) 및 제어 인터페이스(1058)를 통해 사용자와 통신할 수 있다. 디스플레이(1054)는, 예를 들어, TFT(Thin-Film-Transistor Liquid Crystal Display) 디스플레이 또는 OLED(Organic Light Emitting Diode) 디스플레이, 또는 다른 적합한 디스플레이 기술일 수 있다. 디스플레이 인터페이스(1056)는 그래픽 및 다른 정보를 사용자에게 제시하기 위해 디스플레이(1054)를 구동시키기 위한 적합한 회로부를 포함할 수 있다. 제어 인터페이스(1058)는 사용자로부터 커맨드들을 수신할 수 있고 이들을 프로세서(1052)에 제출하기 위해 변환할 수 있다. 더욱이, 디바이스(1050)의 다른 디바이스들과의 근거리 통신을 가능하게 하도록, 프로세서(1052)와 통신하는 외부 인터페이스(1062)가 제공될 수 있다. 외부 인터페이스(1062)는, 예를 들어, 일부 구현예들에서 유선 통신을 위해, 또는 다른 구현예들에서 무선

통신을 위해 제공될 수 있고, 다수의 인터페이스들이 또한 사용될 수 있다.

[0142] 메모리(1064)는 컴퓨팅 디바이스(1050) 내에 정보를 저장한다. 메모리(1064)는 컴퓨터 판독가능 매체 또는 매체들, 휘발성 메모리 유닛 또는 유닛들, 또는 비휘발성 메모리 유닛 또는 유닛들 중 하나 이상으로서 구현될 수 있다. 확장 메모리(1074)는 또한, 예를 들어, SIMM(Single In Line Memory Module) 카드 인터페이스를 포함할 수 있는 확장 인터페이스(1072)를 통해 디바이스(1050)에 제공되고 접속될 수 있다. 그러한 확장 메모리(1074)는 디바이스(1050)에 대한 여분의 저장 공간을 제공할 수 있거나, 디바이스(1050)에 대한 애플리케이션들 또는 다른 정보를 또한 저장할 수 있다. 구체적으로, 확장 메모리(1074)는 전술된 프로세스들을 수행하거나 보완하기 위한 명령어들을 포함할 수 있고, 보안 정보를 또한 포함할 수 있다. 따라서, 예를 들어, 확장 메모리(1074)는 디바이스(1050)를 위한 보안 모듈로서 제공될 수 있고, 디바이스(1050)의 보안 사용을 허용하는 명령어들로 프로그래밍될 수 있다. 더욱이, 보안 애플리케이션들은, 추가 정보와 함께, SIMM 카드들 - 예컨대, 식별 정보를 SIMM 카드에 해킹불가능한 방식으로 둠 - 을 통해 제공될 수 있다.

[0143] 메모리는, 하기에 논의되는 바와 같이, 예를 들어 플래시 메모리 및/또는 NVRAM 메모리를 포함할 수 있다. 일 구현예에서, 컴퓨터 프로그램 제품이 정보 캐리어 내에 유형적으로 구현된다. 컴퓨터 프로그램 제품은, 실행될 때 전술된 것들과 같은 하나 이상의 방법들을 수행하는 명령어들을 포함한다. 정보 캐리어는, 예를 들어 송수신기(1068) 또는 외부 인터페이스(1062)를 통해 수신될 수 있는 메모리(1064), 확장 메모리(1074), 또는 프로세서(1052) 상의 메모리와 같은 컴퓨터 또는 기계 판독가능 매체이다.

[0144] 디바이스(1050)는, 필요한 경우 디지털 신호 처리 회로부를 포함할 수 있는 통신 인터페이스(1066)를 통해 무선으로 통신할 수 있다. 통신 인터페이스(1066)는, 특히, GSM 음성 통화, SMS, EMS, 또는 MMS 메시징, CDMA, TDMA, PDC, WCDMA, CDMA2000, 또는 GPRS와 같은 다양한 모드들 또는 프로토콜들 하에서의 통신을 제공할 수 있다. 그러한 통신은, 예를 들어 무선 주파수 송수신기(1068)를 통해, 일어날 수 있다. 더욱이, 단거리 통신은 예컨대 블루투스, Wi-Fi, 또는 다른 그러한 송수신기(도시되지 않음)를 사용하여 일어날 수 있다. 더욱이, GPS(Global Positioning System) 수신기 모듈(1070)이 추가적인 내비게이션 및 위치 관련 무선 데이터를 디바이스(1050)에 제공할 수 있고, 이는 디바이스(1050) 상에서 실행되는 애플리케이션들에 의해 적절하게 사용될 수 있다.

[0145] 디바이스(1050)는 또한, 오디오 코덱(1060)을 사용하여 청각적으로 통신할 수 있는데, 이는 사용자로부터 구두 정보(spoken information)를 수신할 수 있고 그것을 사용가능한 디지털 정보로 변환할 수 있다. 오디오 코덱(1060)은 마찬가지로, 예컨대 디바이스(1050)의 핸드셋(handset)에서, 예컨대 스피커를 통해, 사용자를 위한 가청 사운드를 생성할 수 있다. 그러한 사운드는 음성 전화 통화들로부터의 사운드를 포함할 수 있고, 기록된 사운드, 예컨대 음성 메시지들, 음악 파일들 등을 포함할 수 있고, 또한, 디바이스(1050) 상에서 동작하는 애플리케이션들에 의해 생성된 사운드를 포함할 수 있다.

[0146] 컴퓨팅 디바이스(1050)는, 도면에 도시된 바와 같이, 다수의 상이한 형태들로 구현될 수 있다. 예를 들어, 그것은 셀룰러 전화기(1080)로서 구현될 수 있다. 그것은 또한, 스마트폰(1082), 개인 휴대 정보 단말기, 또는 다른 유사한 모바일 디바이스의 일부로서 구현될 수 있다.

[0147] 본 명세서에서 설명되는 시스템들 및 방법들의 다양한 구현예들은 디지털 전자 회로부, 집적 회로부, 특수 설계된 ASIC(application specific integrated circuit), 컴퓨터 하드웨어, 펌웨어, 소프트웨어, 및/또는 그러한 구현예들의 조합들로 실현될 수 있다. 이들 다양한 구현예들은, 저장 시스템, 적어도 하나의 입력 디바이스, 및 적어도 하나의 출력 디바이스로부터 데이터 및 명령어들을 수신하고 그들에 데이터 및 명령어들을 송신하도록 결합된, 전용 또는 범용일 수 있는, 적어도 하나의 프로그래밍가능 프로세서를 포함하는 프로그래밍가능 시스템 상에서 실행가능한 그리고/또는 해석가능한 하나 이상의 컴퓨터 프로그램들에서의 구현예를 포함할 수 있다.

[0148] 이러한 컴퓨터 프로그램들(프로그램들, 소프트웨어, 소프트웨어 애플리케이션들 또는 코드로도 알려짐)은 프로그래밍가능 프로세서에 대한 기계 명령어들을 포함하고, 하이-레벨 절차 및/또는 객체 지향 프로그래밍 언어로, 그리고/또는 어셈블리/기계 언어로 구현될 수 있다. 본 명세서에 사용되는 바와 같이, 용어 "기계 판독가능 매체", "컴퓨터 판독가능 매체"는, 기계 판독가능 신호로서 기계 명령어들을 수신하는 기계 판독가능 매체를 포함하는, 프로그래밍가능 프로세서에 기계 명령어들 및/또는 데이터를 제공하는 데 사용되는, 임의의 컴퓨터 프로그램 제품, 장치 및/또는 디바이스, 예를 들어 자기 디스크, 광 디스크, 메모리, PLD(Programmable Logic Device)를 지칭한다. 용어 "기계 판독가능 신호"는 기계 명령어들 및/또는 데이터를 프로그래밍가능 프로세서에 제공하는 데 사용되는 임의의 신호를 지칭한다.

[0149] 사용자와의 상호작용을 제공하기 위해, 본 명세서에서 설명되는 시스템들 및 기법들은 디스플레이 디바이스, 예컨대, 사용자에게 정보를 디스플레이하기 위한 CRT(cathode ray tube) 또는 LCD(liquid crystal display) 모니터, 및 사용자가 컴퓨터에 입력을 제공할 수 있게 하는 키보드 및 포인팅 디바이스, 예컨대 마우스 또는 트랙볼을 갖는 컴퓨터 상에서 구현될 수 있다. 다른 종류의 디바이스들이 또한, 사용자와의 상호작용을 제공하기 위해 사용될 수 있는데; 예를 들어, 사용자에게 제공된 피드백은 임의의 형태의 감각 피드백, 예컨대 시각적 피드백, 청각적 피드백, 또는 촉각적 피드백일 수 있고; 사용자로부터의 입력은 음향, 스피치, 또는 촉각적 입력을 포함하는 임의의 형태로 수신될 수 있다.

[0150] 본 명세서에서 설명되는 시스템들 및 기법들은, 예컨대 데이터 서버로서 백 엔드(back end) 컴포넌트를 포함하는, 또는 미들웨어 컴포넌트, 예컨대 애플리케이션 서버를 포함하는, 또는 프론트 엔드(front end) 컴포넌트, 예컨대 그래픽 사용자 인터페이스 또는 사용자가 본 명세서에서 설명되는 시스템들 및 기법들의 구현과 상호작용할 수 있게 하는 웹 브라우저를 갖는 클라이언트 컴퓨터를 포함하는, 또는 그러한 백 엔드, 미들웨어, 또는 프론트 엔드 컴포넌트들의 임의의 조합을 포함하는 컴퓨팅 시스템에서 구현될 수 있다. 시스템의 컴포넌트들은 디지털 데이터 통신의 임의의 형태 또는 매체, 예컨대 통신 네트워크에 의해 상호접속될 수 있다. 통신 네트워크들의 예들은 근거리 네트워크("LAN"), 광역 네트워크("WAN"), 및 인터넷을 포함한다.

[0151] 컴퓨팅 시스템은 클라이언트들 및 서버들을 포함할 수 있다. 클라이언트 및 서버는 대체적으로 서로 원격이며, 전형적으로 통신 네트워크를 통해 상호작용한다. 클라이언트와 서버의 관계는, 각각의 컴퓨터들 상에서 실행되고 서로에 대해 클라이언트-서버 관계를 갖는 컴퓨터 프로그램들에 의해 발생한다.

[0152] 다수의 실시예들이 기술되었다. 그럼에도 불구하고, 본 발명의 사상 및 범주로부터 벗어남이 없이 다양한 변경들이 이루어질 수 있다는 것이 이해될 것이다. 더욱이, 도면들에 묘사된 논리 흐름들은 바람직한 결과들을 달성하기 위해 도시된 특정 순서, 또는 순차적 순서를 요구하지 않는다. 더욱이, 다른 단계들이 제공될 수 있거나, 설명된 흐름들로부터 단계들이 제거될 수 있고, 다른 컴포넌트들이 설명된 시스템들에 추가되거나 그로부터 제거될 수 있다. 따라서, 다른 실시예들이 하기의 청구범위의 범주 내에 있다.

[0153] 본 명세서에 설명된 본 발명의 실시예들 및 모든 기능 동작들은, 디지털 전자 회로부에서, 또는 본 명세서에서 개시되는 구조들 및 이들의 구조적 등가물을 포함하는 컴퓨터 소프트웨어, 펌웨어 또는 하드웨어에서, 또는 이들 중 하나 이상의 조합들로 구현될 수 있다. 본 발명의 실시예들은 하나 이상의 컴퓨터 프로그램 제품들, 예를 들어, 데이터 처리 장치에 의한 실행을 위해 또는 데이터 처리 장치의 동작을 제어하기 위해 컴퓨터 판독 가능 매체 상에 인코딩된 컴퓨터 프로그램 명령어들의 하나 이상의 모듈들로서 구현될 수 있다. 컴퓨터 판독가능 매체는 기계 판독가능 저장 디바이스, 기계 판독가능 저장 기판, 메모리 디바이스, 기계 판독가능 전파 신호에 영향을 미치는 물질의 조성물, 또는 이들 중 하나 이상의 조합일 수 있다. 용어 "데이터 처리 장치"는 예를 들어 프로그래밍가능 프로세서, 컴퓨터, 또는 다수의 프로세서들 또는 컴퓨터들을 포함하는, 데이터를 처리하기 위한 모든 장치, 디바이스, 및 기계를 포함한다. 장치는, 하드웨어에 더하여, 해당 컴퓨터 프로그램을 위한 실행 환경을 생성하는 코드, 예를 들어 프로세서 펌웨어, 프로토콜 스택, 데이터베이스 관리 시스템, 운영 체제 또는 이들의 하나 이상의 조합을 구성하는 코드를 포함할 수 있다. 전파 신호는 인위적으로 생성된 신호, 예를 들어, 적합한 수신기 장치로의 송신을 위한 정보를 인코딩하기 위해 생성되는 기계 생성된 전기, 광학 또는 전자기 신호이다.

[0154] 컴퓨터 프로그램(또한, 프로그램, 소프트웨어, 소프트웨어 애플리케이션, 스크립트 또는 코드로서 알려짐)은 컴파일된 또는 해석된 언어를 포함하는 임의의 형태의 프로그래밍 언어로 기록될 수 있고, 이는 독립형 프로그램으로서, 또는 모듈, 컴포넌트, 서브루틴, 또는 컴퓨팅 환경에서 사용하기에 적합한 다른 유닛으로서의 형태를 포함하는 임의의 형태로 배포(deploy)될 수 있다. 컴퓨터 프로그램은 반드시 파일 시스템 내의 파일에 대응하는 것은 아니다. 프로그램은 다른 프로그램들 또는 데이터(예로서, 마크업 언어 문서 내에 저장된 하나 이상의 스크립트들)를 유지하는 파일의 일부 내에, 해당 프로그램에 전용화된 단일 파일 내에, 또는 다수의 협력 파일들(예를 들어, 하나 이상의 모듈들, 서브프로그램들, 또는 코드의 부분들을 저장하는 파일들) 내에 저장될 수 있다. 컴퓨터 프로그램은 하나의 컴퓨터 상에서, 또는 하나의 사이트에 위치되거나 다수의 사이트들에 걸쳐 분산되고 통신 네트워크에 의해 상호접속되는 다수의 컴퓨터들 상에서 실행되도록 배치될 수 있다.

[0155] 본 명세서에서 설명된 프로세스들, 동작들, 및/또는 논리 흐름들은 본 명세서에 설명된 프로세스들, 동작들, 및/또는 논리 흐름들의 기능을 실현하기 위해 하나 이상의 컴퓨터 소프트웨어 명령어들을 실행하는 하나 이상의 중앙 처리 장치(CPU)들 또는 그래픽 처리 장치(GPU)들에 의해 수행될 수 있다. 프로세스들, 동작들, 및/또는 논리 흐름들은 또한 하드웨어 회로부에서 수행될 수 있다. 예를 들어, 일부 구현예들에서, 본 개시내용의 동작들

은 본 명세서에 설명된 프로세스들, 동작들 및/또는 논리 흐름들의 기능을 실현하도록 프로그래밍식으로 구성된 FPGA(field programmable gate array)의 로직 게이트들을 사용하여 구현된 처리 엔진들에 의해 실행될 수 있다. 다른 예로서, 본 개시내용의 동작들은 본 명세서에 설명된 프로세스들, 동작들, 및/또는 논리 흐름들의 기능을 실현하도록 구성된 주문형 집적 회로(ASIC)의 로직 게이트들을 사용하여 구현된 처리 엔진들에 의해 실행될 수 있다. 또 다른 구현예들에서, 프로세스들, 동작들, 및/또는 논리 흐름들의 부분들의 일부는 하나 이상의 CPU들 또는 하나 이상의 GPU들에 의해 구현될 수 있고, 프로세스들, 동작들, 및/또는 논리 흐름들의 부분들의 일부는 하드웨어 회로부에 의해 그리고 임의의 순서로 구현될 수 있다.

[0156] 컴퓨터 프로그램의 실행에 적절한 프로세서들은, 예를 들어 범용 및 전용 둘 모두의 마이크로프로세서들, 및 임의의 종류의 디지털 컴퓨터의 임의의 하나 이상의 프로세서들을 포함한다. 일반적으로, 프로세서는 판독 전용 메모리 또는 랜덤 액세스 메모리 또는 둘 모두로부터 명령어들 및 데이터를 수신할 것이다. 컴퓨터의 본질적인 요소들은 명령어들을 수행하기 위한 프로세서, 및 명령어들 및 데이터를 저장하기 위한 하나 이상의 메모리 디바이스들이다. 일반적으로, 컴퓨터는 또한 데이터를 저장하기 위한 하나 이상의 대용량 저장 디바이스들, 예를 들어 자기, 광자기 디스크 또는 광 디스크를 포함하거나, 또는 그들로부터 데이터를 수신하거나, 그들에 데이터를 전달하거나, 둘 모두를 행하도록 동작가능하게 결합될 것이다. 그러나, 컴퓨터는 그러한 디바이스들을 가질 필요는 없다. 또한, 컴퓨터는 다른 디바이스에, 예를 들어, 몇 가지 예를 들면, 태블릿 컴퓨터, 휴대 전화, PDA(Personal Digital Assistant), 모바일 오디오 재생기, GPS(Global Positioning System) 수신기에 내장될 수 있다. 컴퓨터 프로그램 명령어들 및 데이터를 저장하는 데 적합한 컴퓨터 판독가능 매체들은 예로서 반도체 메모리 디바이스, 예를 들어, EPROM, EEPROM 및 플래시 메모리 디바이스; 자기 디스크, 예를 들어, 내부 하드 디스크 또는 이동식 디스크; 자기광학 디스크; 및 CDROM 및 DVD-ROM 디스크를 포함하는 모든 형태의 비휘발성 메모리, 매체 및 메모리 디바이스를 포함한다. 프로세서 및 메모리는 특수 목적 논리 회로부에 의해 보완되거나 그 안에 통합될 수 있다.

[0157] 사용자와의 상호작용을 제공하기 위해, 본 발명의 실시예들은 디스플레이 디바이스, 예컨대, 사용자에게 정보를 디스플레이하기 위한 CRT(cathode ray tube) 또는 LCD(liquid crystal display) 모니터, 및 사용자가 컴퓨터에 입력을 제공할 수 있게 하는 키보드 및 포인팅 디바이스, 예컨대 마우스 또는 트랙볼을 갖는 컴퓨터 상에서 구현될 수 있다. 다른 종류의 디바이스들이 또한, 사용자와의 상호작용을 제공하기 위해 사용될 수 있는데; 예를 들어, 사용자에게 제공된 피드백은 임의의 형태의 감각 피드백, 예컨대 시각적 피드백, 청각적 피드백, 또는 촉각적 피드백일 수 있고; 사용자로부터의 입력은 음향, 스피치, 또는 촉각적 입력을 포함하는 임의의 형태로 수신될 수 있다.

[0158] 본 발명의 실시예들은, 예를 들어 데이터 서버로서 백엔드 컴포넌트를 포함하거나, 또는 미들웨어 컴포넌트, 예를 들어 응용 서버를 포함하거나, 또는 프론트엔드 컴포넌트, 예를 들어 사용자가 본 발명의 구현예와 상호작용할 수 있게 하는 웹 브라우저 또는 그래픽 사용자 인터페이스를 갖는 클라이언트 컴퓨터를 포함하거나, 또는 하나 이상의 이러한 백엔드, 미들웨어, 또는 프론트엔드 컴포넌트들의 임의의 조합을 포함하는 컴퓨팅 시스템에서 구현될 수 있다. 시스템의 컴포넌트들은 디지털 데이터 통신의 임의의 형태 또는 매체, 예컨대 통신 네트워크에 의해 상호접속될 수 있다. 통신 네트워크들의 예들은 근거리 네트워크("LAN") 및 광역 네트워크("WAN"), 예컨대, 인터넷을 포함한다.

[0159] 컴퓨팅 시스템은 클라이언트들 및 서버들을 포함할 수 있다. 클라이언트 및 서버는 대체적으로 서로 원격이며, 전형적으로 통신 네트워크를 통해 상호작용한다. 클라이언트와 서버의 관계는 각자의 컴퓨터들 상에서 실행되고 서로 클라이언트-서버 관계를 갖는 컴퓨터 프로그램들에 의해 발생한다.

[0160] 본 명세서는 많은 세부사항을 포함하지만, 이들은 본 발명 또는 청구될 수 있는 것의 범주에 대한 제한으로서가 아니라, 오히려 본 발명의 특정 실시예들에 특정적인 특징들의 설명으로서 해석되어야 한다. 별개의 실시예들과 관련하여 본 명세서에 설명된 특정 특징들은 또한 단일 실시예에서 조합되어 구현될 수 있다. 반대로, 단일 실시예와 관련하여 설명된 다양한 특징들은 또한 다수의 실시예들에서 별개로 또는 임의의 적합한 하위조합으로 구현될 수 있다. 더욱이, 위에서 특징들은 특정 조합들에서 동작하는 것으로 설명되고 심지어 처음에 그와 같이 청구될 수 있지만, 청구되는 조합으로부터의 하나 이상의 특징들은 일부 경우들에서 조합으로부터 제거될 수 있으며, 청구되는 조합은 하위조합 또는 하위조합의 변형과 관련될 수 있다.

[0161] 유사하게, 도면들에는 동작들이 특정 순서로 도시되지만, 이것은 바람직한 결과들을 달성하기 위해 그러한 동작들이 도시된 특정 순서로 또는 순차적인 순서로 수행되어야 하거나, 모든 예시된 동작들이 수행되어야 하는 것을 요구하는 것으로 이해되지 않아야 한다. 특정 상황들에서는, 다중작업 및 병렬 처리가 유리할 수 있다. 더욱

이, 전술된 실시예들에서의 다양한 시스템 컴포넌트들의 분리는 모든 실시예들에서 그러한 분리를 요구하는 것으로 이해되어서는 안되며, 설명된 프로그램 컴포넌트들 및 시스템들은 일반적으로 단일 소프트웨어 제품에 함께 통합되거나 다수의 소프트웨어 제품들로 패키징될 수 있음을 이해하여야 한다.

[0162] HTML 파일이 언급되는 각각의 경우에서, 다른 파일 유형들 또는 포맷들이 대체될 수 있다. 예를 들어, HTML 파일은 XML, JSON, 일반 텍스트, 또는 다른 유형의 파일들에 의해 대체될 수 있다. 또한, 테이블 또는 해시 테이블이 언급되는 경우, 다른 데이터 구조들(예를 들어, 스프레드시트, 관계형 데이터베이스, 또는 구조화된 파일)이 사용될 수 있다.

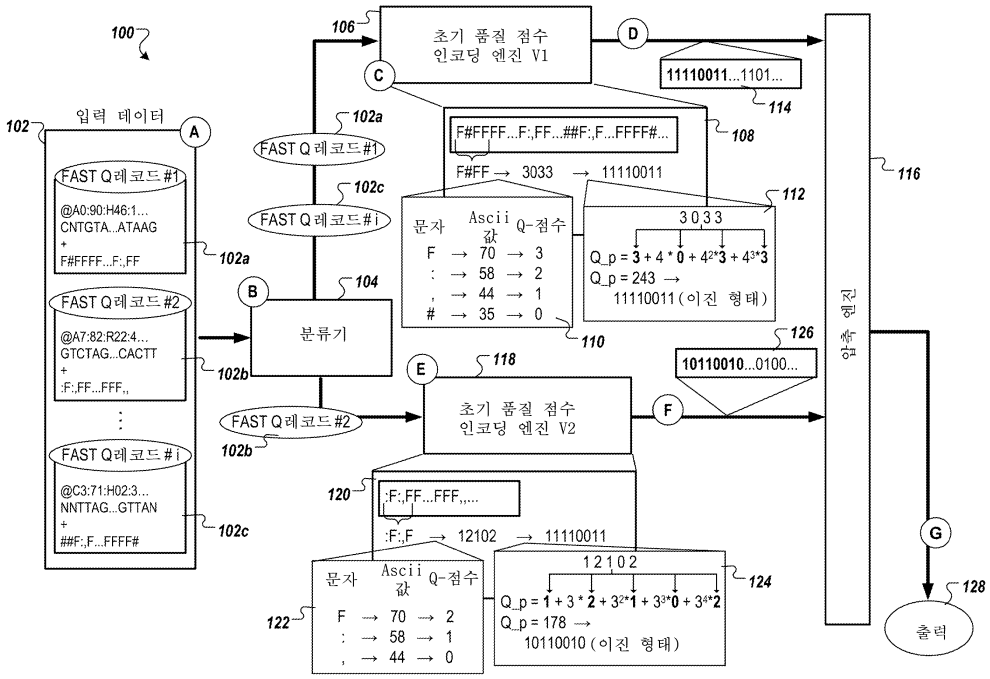
[0163] 본 발명의 특정 실시예들이 설명되었다. 다른 실시예들이 하기의 청구범위의 범주 내에 있다. 예를 들어, 청구범위에 열거된 단계들은 상이한 순서로 수행될 수 있고, 여전히 바람직한 결과들을 달성할 수 있다.

[0164] **다른 실시예들**

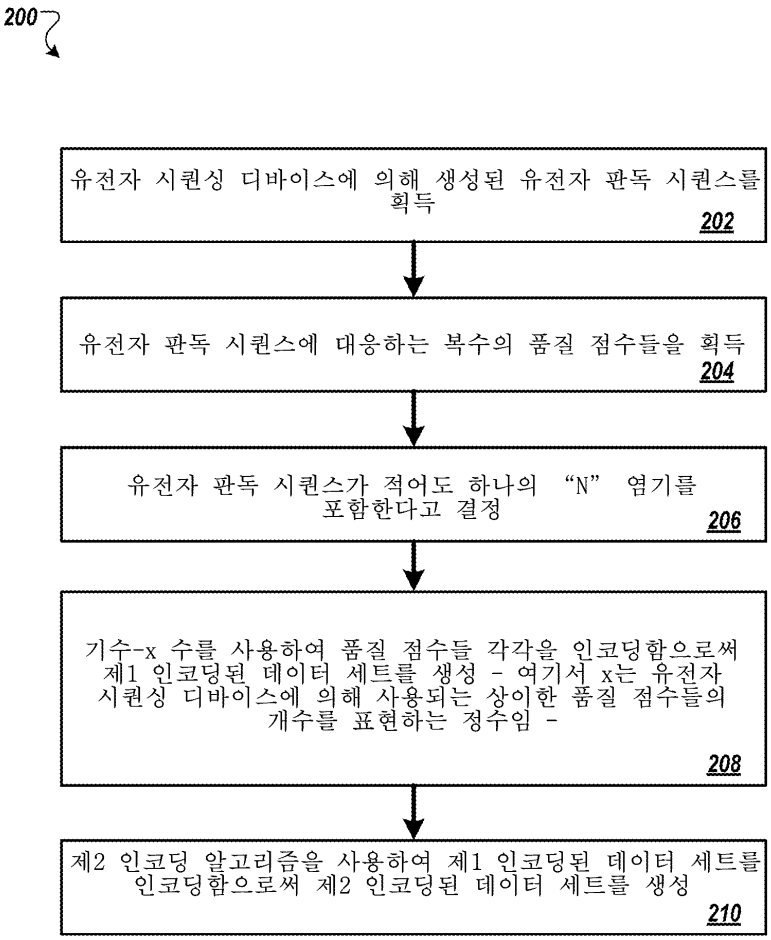
[0165] 본 발명의 특정 실시예들이 설명되었다. 다른 실시예들이 하기의 청구범위의 범주 내에 있다. 예를 들어, 청구범위에 열거된 단계들은 상이한 순서로 수행될 수 있고, 여전히 바람직한 결과들을 달성할 수 있다.

도면

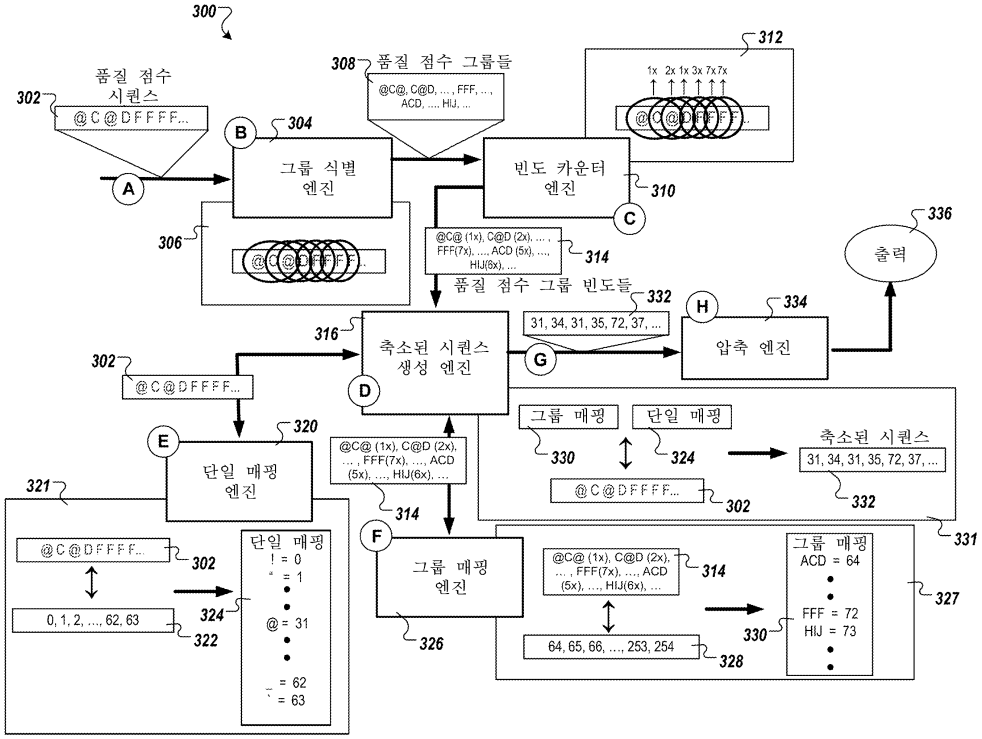
도면1



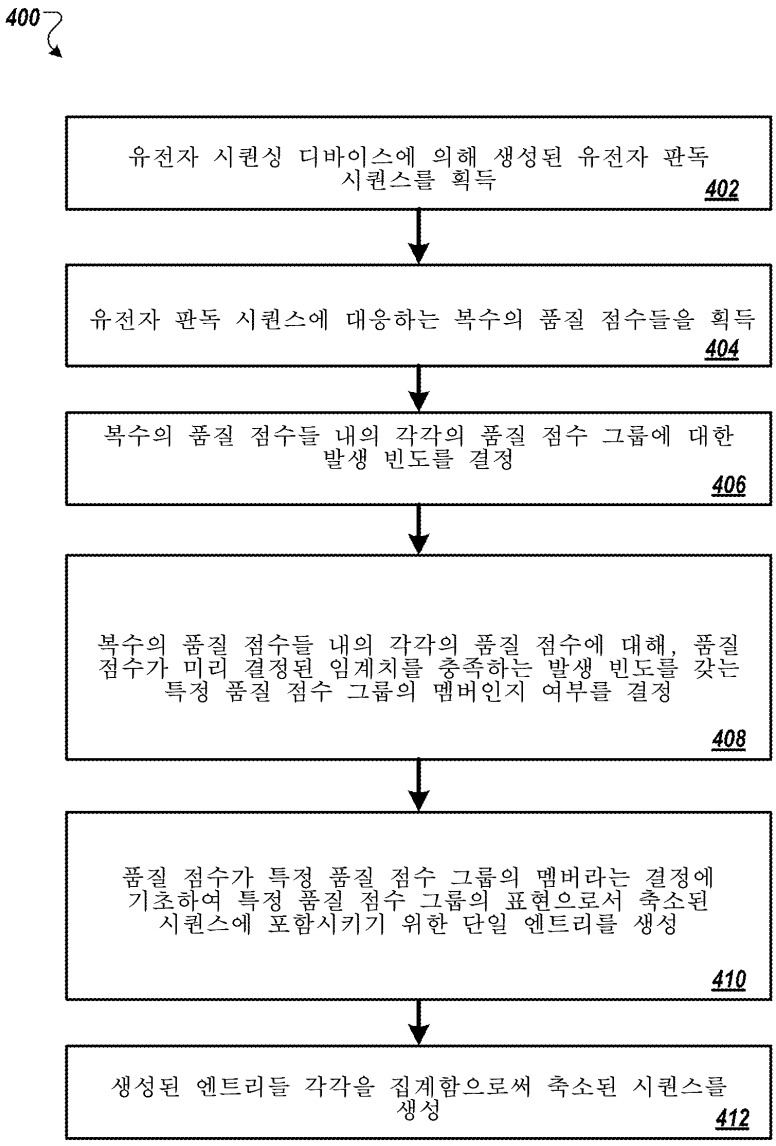
도면2


200
유전자 시퀀싱 디바이스에 의해 생성된 유전자 판독 시퀀스를 획득
202
유전자 판독 시퀀스에 대응하는 복수의 품질 점수들을 획득
204
유전자 판독 시퀀스가 적어도 하나의 “N” 염기를 포함한다고 결정
206
기수-x 수를 사용하여 품질 점수들 각각을 인코딩함으로써 제1 인코딩된 데이터 세트를 생성 - 여기서 x는 유전자 시퀀싱 디바이스에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수임 -
208
제2 인코딩 알고리즘을 사용하여 제1 인코딩된 데이터 세트를 인코딩함으로써 제2 인코딩된 데이터 세트를 생성
210

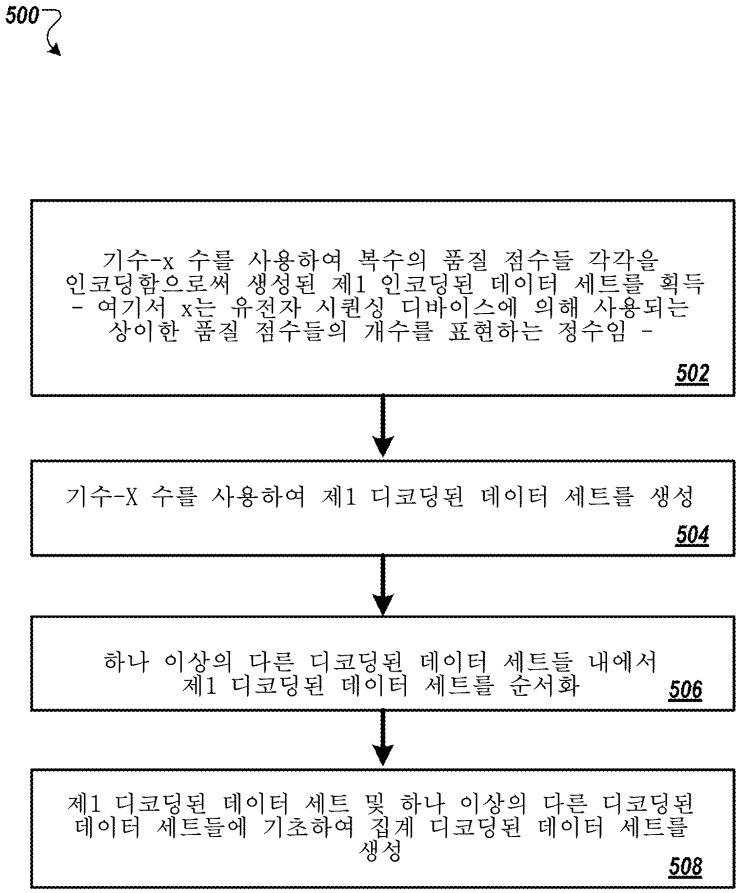
도면3


300
품질 점수 시퀀스
302
@C@DFFFF...
A
B
304
그룹 식별 엔진
306
308
품질 점수 그룹들
@C@, C@D, ..., FFF, ..., ACD, ..., HIJ, ...
312
1x 2x 1x 3x 7x7x
빈도 카운터 엔진
310
C
@C@ (1x), C@D (2x), ..., FFF(7x), ..., ACD (5x), ..., HIJ(6x), ...
314
품질 점수 그룹 빈도들
332
31, 34, 31, 35, 72, 37, ...
H
334
압축 엔진
336
출력
316
축소된 시퀀스 생성 엔진
D
G
E
320
단일 매핑 엔진
321
F
그룹 매핑 엔진
326
그룹 매핑
단일 매핑
330
324
축소된 시퀀스
331
327
0, 1, 2, ..., 62, 63
322
! = 0
" = 1
@ = 31
_ = 62
` = 63
ACD = 64
FFF = 72
HIJ = 73
64, 65, 66, ..., 253, 254
328

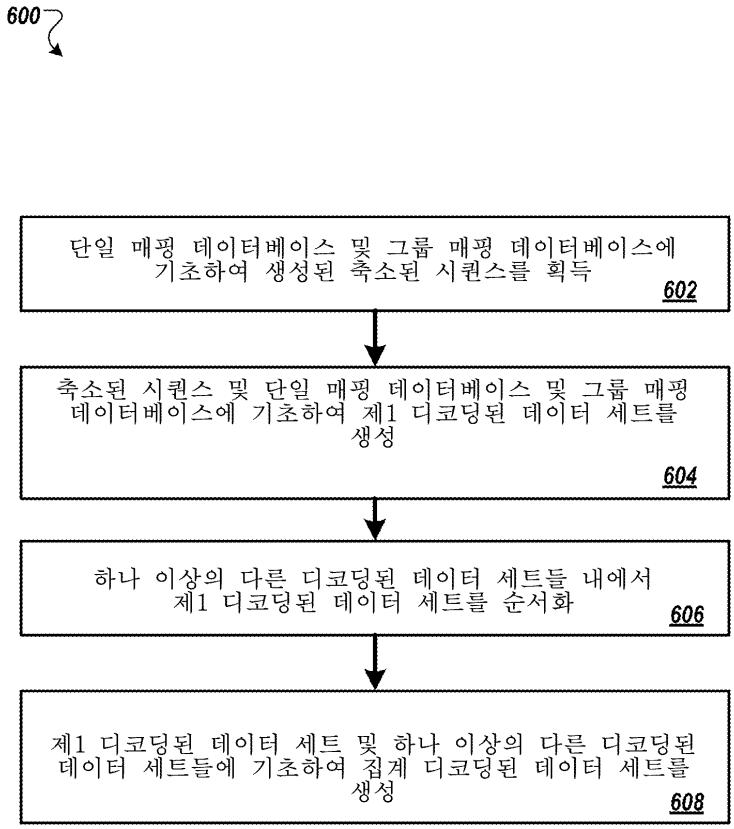
도면4


400
유전자 시퀀싱 디바이스에 의해 생성된 유전자 판독 시퀀스를 획득
402
유전자 판독 시퀀스에 대응하는 복수의 품질 점수들을 획득
404
복수의 품질 점수들 내의 각각의 품질 점수 그룹에 대한 발생 빈도를 결정
406
복수의 품질 점수들 내의 각각의 품질 점수에 대해, 품질 점수가 미리 결정된 임계치를 충족하는 발생 빈도를 갖는 특정 품질 점수 그룹의 멤버인지 여부를 결정
408
품질 점수가 특정 품질 점수 그룹의 멤버라는 결정에 기초하여 특정 품질 점수 그룹의 표현으로서 축소된 시퀀스에 포함시키기 위한 단일 엔트리를 생성
410
생성된 엔트리들 각각을 집계함으로써 축소된 시퀀스를 생성
412

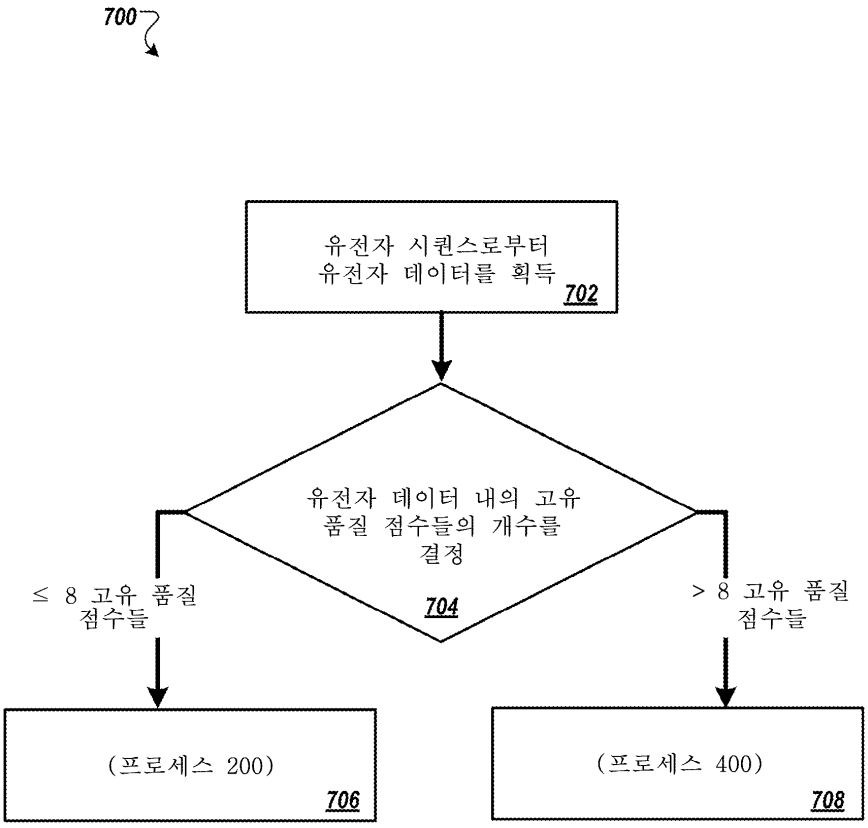
도면5


500
기수-x 수를 사용하여 복수의 품질 점수들 각각을 인코딩함으로써 생성된 제1 인코딩된 데이터 세트를 획득 - 여기서 x는 유전자 시퀀싱 디바이스에 의해 사용되는 상이한 품질 점수들의 개수를 표현하는 정수임 -
502
기수-X 수를 사용하여 제1 디코딩된 데이터 세트를 생성
504
하나 이상의 다른 디코딩된 데이터 세트들 내에서 제1 디코딩된 데이터 세트를 순서화
506
제1 디코딩된 데이터 세트 및 하나 이상의 다른 디코딩된 데이터 세트들에 기초하여 집계 디코딩된 데이터 세트를 생성
508

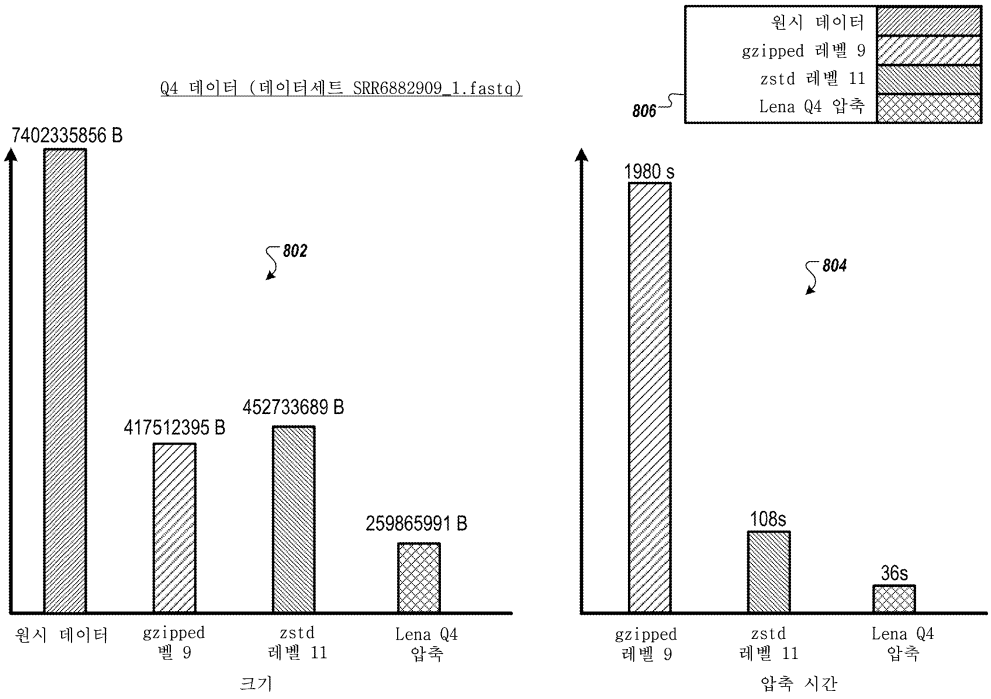
도면6


600
단일 매핑 데이터베이스 및 그룹 매핑 데이터베이스에 기초하여 생성된 축소된 시퀀스를 획득
602
축소된 시퀀스 및 단일 매핑 데이터베이스 및 그룹 매핑 데이터베이스에 기초하여 제1 디코딩된 데이터 세트를 생성
604
하나 이상의 다른 디코딩된 데이터 세트들 내에서 제1 디코딩된 데이터 세트를 순서화
606
제1 디코딩된 데이터 세트 및 하나 이상의 다른 디코딩된 데이터 세트들에 기초하여 집계 디코딩된 데이터 세트를 생성
608

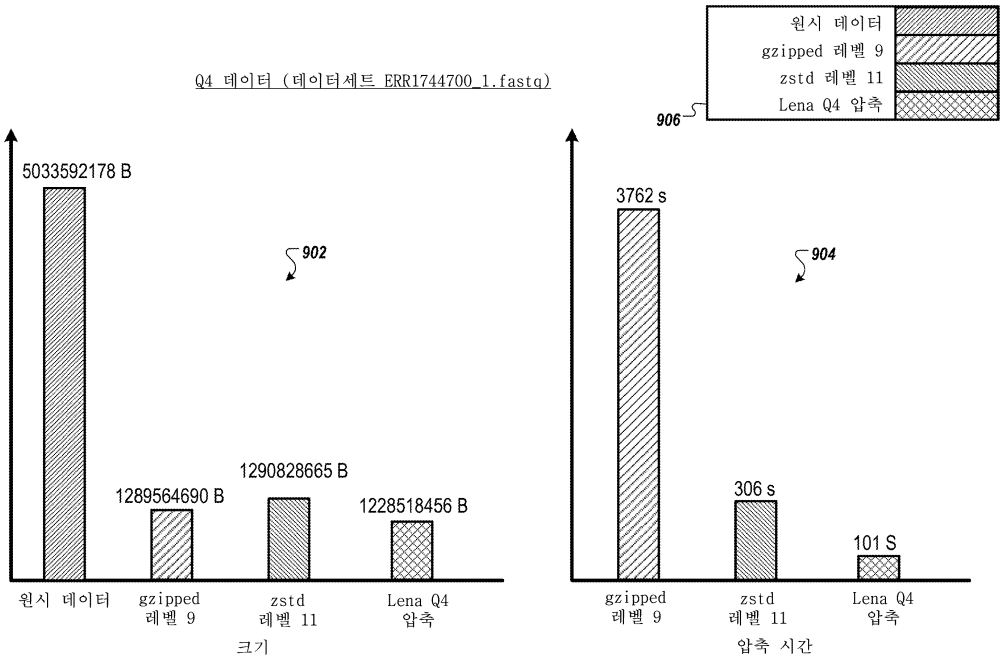
도면7


700
유전자 시퀀스로부터
유전자 데이터를 획득
702
유전자 데이터 내의 고유
품질 점수들의 개수를
결정
704
≤ 8 고유 품질
점수들
> 8 고유 품질
점수들
(프로세스 200)
706
(프로세스 400)
708

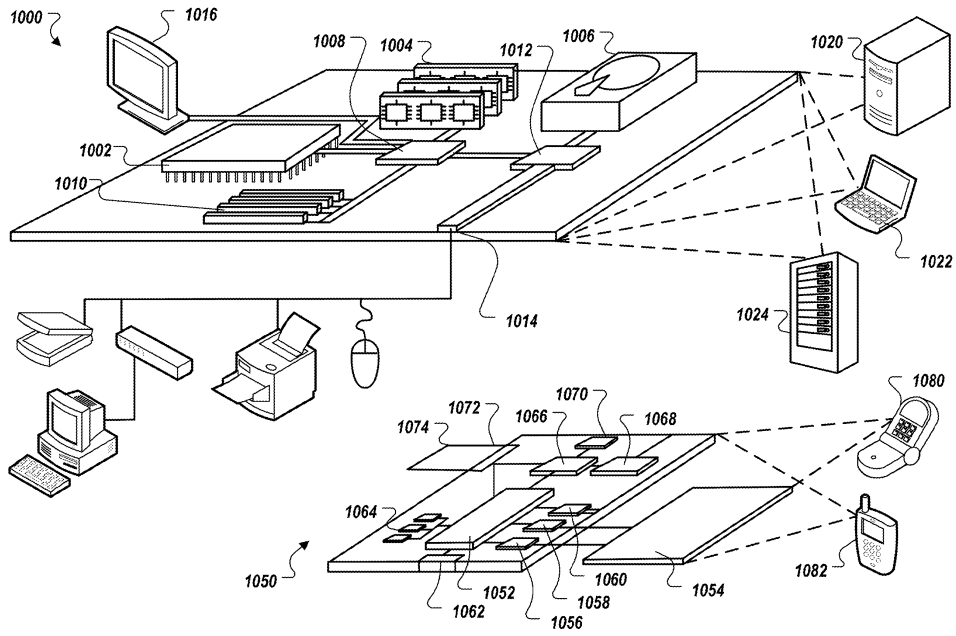
도면8


Q4 데이터 (데이터세트 SRR6882909_1.fastq)
원시 데이터
gzipped 레벨 9
zstd 레벨 11
Lena Q4 압축
806
7402335856 B
417512395 B
452733689 B
259865991 B
802
원시 데이터
gzipped
벨 9
zstd
레벨 11
Lena Q4
압축
크기
1980 s
108s
36s
804
gzipped
레벨 9
zstd
레벨 11
Lena Q4
압축
압축 시간

도면9


Q4 데이터 (데이터세트 ERR1744700_1.fastq)
원시 데이터
gzipped 레벨 9
zstd 레벨 11
Lena Q4 압축
906
5033592178 B
1289564690 B
1290828665 B
1228518456 B
902
원시 데이터
gzipped 레벨 9
zstd 레벨 11
Lena Q4 압축
크기
3762 s
306 s
101 S
904
gzipped 레벨 9
zstd 레벨 11
Lena Q4 압축
압축 시간

도면10


1000
1016
1008
1004
1012
1006
1020
1002
1010
1022
1014
1024
1080
1070
1072
1066
1068
1074
1064
1050
1052
1062
1060
1058
1056
1054
1082