(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년05월11일
(11) 등록번호 10-2396789
(24) 등록일자 2022년05월06일

(51) 국제특허분류(Int. Cl.)
G16B 50/50 (2019.01) *G16B 30/10* (2019.01)
G16B 30/20 (2019.01) *H03M 7/30* (2006.01)
(52) CPC특허분류
G16B 50/50 (2019.02)
G16B 30/10 (2019.02)
(21) 출원번호 10-2019-0143734
(22) 출원일자 2019년11월11일
심사청구일자 2019년11월11일
(65) 공개번호 10-2021-0056822
(43) 공개일자 2021년05월20일
(56) 선행기술조사문헌
Y. Xing 외, "GTZ: a fast compression and cloud transmission tool optimized for FASTQ files", BMC Bioinformatics 18(Suppl 16), 2017.*
KR1020180086484 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
주식회사 셀젠텍
충청북도 청주시 흥덕구 오송읍 오송생명 2로 110-6
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김희율
대전광역시 유성구 유성대로 579-4 (구암동)
김동우
충청북도 청주시 청원구 오창읍 오창중앙로 27 코아루아파트 301동1404호
(뒷면에 계속)
(74) 대리인
민병준

전체 청구항 수 : 총 6 항 심사관 : 성경아

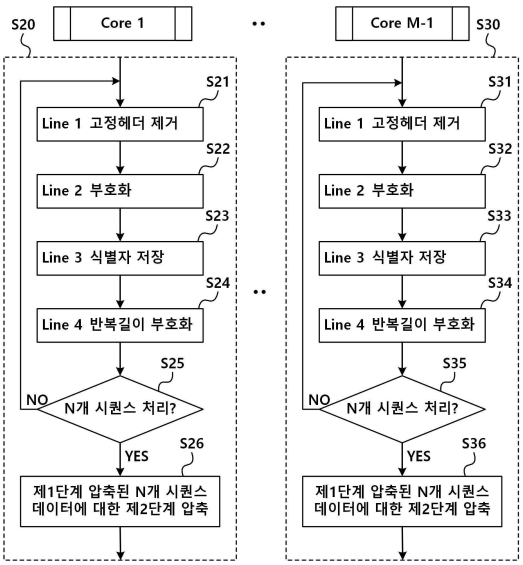
(54) 발명의 명칭 **FASTQ 포맷의 유전체 데이터를 위한 유전체 데이터의 압축 및 전송 방법**

(57) 요 약

본 발명의 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법은, M개의 코어중 하나의 코어인 제 1 코어가 첫번째 시퀀스 데이터의 첫번째 라인에서 고정 헤더 데이터를 압축결과 저장소에 저장하는 단계; 상기 제 1 코어가 나머지 M-1(M은 4이상의 자연수)개의 코어(이하 '기타 코어들'이라 한다) 각각에 대하여 N(N은 2

(뒷면에 계속)

대 표 도 - 도3



이상의 자연수)개씩의 시퀀스 데이터를 분배하고 기타 코어들에서 각각 압축되도록 하여, 한번에 N*(M-1)개의 시퀀스 데이터에 대한 압축을 병렬 처리하고 상기 압축결과 저장소에 저장되도록 하는 단계;를 포함하되, 상기한 기타 코어들의 각각에서 실행되는 압축은, 각 시퀀스 데이터에 대하여, 첫번째 라인의 고정헤더를 제거하는 과정; 두번째 라인을 부호화하는 과정; 세번째 라인의 식별자를 저장하는 과정; 네번째 라인을 반복길이 부호화하는 과정;으로 구성되는 처리 과정을 상기 N개의 시퀀스 데이터에 대하여 반복하는 제 1 단계 압축과, 상기 N개의 시퀀스 데이터에 대한 상기 제1 단계 압축의 결과에 대하여, 무손실 압축 알고리즘에 의해 압축하는 제 2 단계 압축;을 실행하는 것을 특징으로 한다.

(52) CPC특허분류
G16B 30/20 (2019.02)
H03M 7/6011 (2013.01)
H03M 7/6023 (2013.01)
H03M 7/70 (2013.01)

(72) 발명자
오성열
충청북도 청주시 청원구 오창읍 오창중앙로 13 우림필유1차아파트 102동1704호
김영준
서울특별시 강남구 논현로160길 31 청록빌라 202호
이진영
경기도 안산시 단원구 초지1로 78 행복한마을아파트 2002동 404호

이 발명을 지원한 국가연구개발사업

과제고유번호	NRF-2017M3A9A7050614
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	바이오, 의료기술개발(R&D)
연구과제명	대장암 특이적 정밀진단 마커 개발을 위한 다중유전체 데이터 연계 확장 분석
기 여 율	1/2
과제수행기관명	연세대학교 산학협력단
연구기간	2019.01.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	NRF-2017M3A9A7050615
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	바이오, 의료기술개발(R&D)
연구과제명	정밀의료 실현을 위한 통합분석 플랫폼 및 운용시스템 개발
기 여 율	2/2
과제수행기관명	주식회사 셀젠텍
연구기간	2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

M개의 코어로 구성되는 프로세서에서 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법으로서,

M개의 코어중 하나의 코어인 제 1 코어가 첫번째 시퀀스 데이터의 첫번째 라인에서 고정 헤더 데이터를 압축결과 저장소에 저장하는 단계;

상기 제 1 코어가 나머지 M-1(M은 4이상의 자연수)개의 코어(이하 '기타 코어들'이라 한다) 각각에 대하여 N(N은 2이상의 자연수)개씩의 시퀀스 데이터를 분배하고 기타 코어들에서 각각 압축되도록 하여, 한번에 N*(M-1)개의 시퀀스 데이터에 대한 압축을 병렬 처리하고 상기 압축결과 저장소에 저장되도록 하는 단계;를 포함하되,

상기한 기타 코어들의 각각에서 실행되는 압축은,

각 시퀀스 데이터에 대하여, 첫번째 라인의 고정헤더를 제거하는 과정; 두번째 라인을 부호화하는 과정; 세번째 라인의 식별자를 저장하는 과정; 네번째 라인을 반복길이 부호화하는 과정;으로 구성되는 처리 과정을 상기 N개의 시퀀스 데이터에 대하여 반복하는 제 1 단계 압축과,

상기 N개의 시퀀스 데이터에 대한 상기 제1 단계 압축의 결과에 대하여, 무손실 압축 알고리즘에 의해 압축하는 제 2 단계 압축;을 실행하는,

것을 특징으로 하는 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법.

청구항 2

청구항 1에 있어서,

상기 제 2 단계 압축에서는 7z 압축 알고리즘이 적용되는 것을 특징으로 하는 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법.

청구항 3

청구항 1에 있어서,

상기 두번째 라인의 부호화는,

사전 설정되는 맵핑 테이블에 의해 A,T,G,C의 문자를 2bit 코드로 맵핑함으로써 실행되는,

것을 특징으로 하는 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법.

청구항 4

삭제

청구항 5

M개의 코어로 구성되는 프로세서를 구비한 송신처에서 FASTQ 포맷의 유전체 데이터를 압축하여 수신처로 전송하는 유전체 데이터의 압축전송 방법으로서,

상기 송신처에서의 압축은,

M개의 코어중 하나의 코어인 제 1 코어가 나머지 M-1(M은 4이상의 자연수)개의 코어(이하 '기타 코어들'이라 한다) 각각에 대하여 N(N은 2이상의 자연수)개씩의 시퀀스 데이터를 분배하고 상기 기타 코어들에서 압축되도록 하여, 한번에 N*(M-1)개의 시퀀스 데이터에 대한 압축을 병렬 처리하는 단계;를 포함하되,

상기한 기타 코어들의 각각에서 실행되는 압축은,

각 시퀀스 데이터에 대하여, 첫번째 라인의 고정헤더를 제거하는 과정; 두번째 라인을 부호화하는 과정; 세번째

라인의 식별자를 저장하는 과정; 네번째 라인을 반복길이 부호화하는 과정;으로 구성되는 처리 과정을 상기 N개의 시퀀스 데이터에 대하여 반복하는 제 1 단계 압축을 포함하여 실행되도록 하며,

상기한 기타 코어들의 각각에서 실행되는 압축은,

상기 N개의 시퀀스 데이터에 대한 상기 제 1 단계 압축의 결과에 대하여, 무손실 압축 알고리즘에 의해 압축하는 제 2 단계 압축;을 선택적으로 실행하는,

것을 특징으로 하는 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축전송 방법.

청구항 6

청구항 5에 있어서,

상기 제 1 코어가 첫번째 시퀀스 데이터의 첫번째 라인에서 고정 헤더 데이터를 수신처로 전송하는 단계;

상기 송신처에서 유전체 데이터중 S(S는 1이상의 자연수)개의 시퀀스 데이터에 대하여 상기 제 1 단계 압축을 실행하여 상기 수신처로 전송하고 상기 수신처에서 압축해제를 실행하며, 상기 제 2 단계 압축을 실행하여 상기 수신처로 전송하고 상기 수신처에서 상기 제 2 단계 압축에 대응하는 압축해제를 실행하되, 하나 이상의 무손실 압축 알고리즘별로 실행하는 속도 측정용 압축 및 전송을 실행하는 단계;

상기 송신처의 제 1 코어가 상기 속도 측정용 압축 및 전송의 결과에 따라, 상기 S개의 시퀀스 데이터를 제외한 나머지 시퀀스 데이터(이하 '메인 시퀀스 데이터'라고 한다)에 대하여 적용할 상기 제 2 단계 압축의 실시 여부와 상기 제 2 단계 압축에 적용할 무손실 알고리즘을 결정함으로써 메인 시퀀스 데이터에 대한 압축 방식을 결정하는 단계;

상기 메인 시퀀스 데이터에 대하여 상기 결정된 압축 방식에 따라 상기 나머지 코어들에 의해 압축이 실행되도록 하며 압축된 결과가 상기 수신처로 전송되도록 하는,

것을 특징으로 하는 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법.

청구항 7

청구항 6에 있어서,

상기한 메인 시퀀스 데이터에 대한 압축 방식의 결정은,

상기 속도 측정용 압축 및 전송의 결과를 이용하여, 상기 제 1 단계 압축을 실시하고 상기 제 2 단계 압축을 실시하지 않는 압축 전송 형태와, 상기 제 1 단계 압축을 실시하고 적어도 하나 이상의 후보가 되는 무손실 압축 알고리즘에 따라 상기 제 2 단계 압축을 실시하는 압축전송 형태의 각각에 대하여,

상기 송신처 및 수신처에서 압축, 전송 및 압축해제에 소요되는 총 예상 소요시간을 계산하고 총 예상 소요시간이 가장 작은 압축전송 형태를 상기 메인 시퀀스 데이터에 대한 압축 방식으로 결정하는,

것을 특징으로 하는 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 FASTQ 포맷의 유전체 데이터를 압축하여 저장하거나 압축하여 전송하는 유전체 데이터의 압축 및 전송 방법에 관한 것이다.

배 경 기 술

[0002] 유전자는 세포의 기능을 담당하는 최소의 단위, 즉 한 단백질을 전사(translation)하는 최소한의 염기(base) 단위를 말한다. 여기서 염기는 문장을 이루는 글자에 해당하며 A(Adenosine), T(Thymine), G(Guanine), C(Cytosine) 네 종류가 있다.

[0003] 유전자를 이루는 염기 ATGC의 결합 순서를 알아내는 기술을 DNA 시퀀싱이라고 한다. HGP(Human Genome Project)에서 사용한 생어 시퀀싱(Sanger Sequencing)이 일반적이었으나, 현재는 NGS(Next Generation Sequencing)가 일반적이며, Single Molecule 시퀀싱, Nanopore 시퀀싱 등 3세대 시퀀싱 방법론들도 제안되고 있다.

[0004] 개인 유전정보 분석 종류에는 WGS(Whole Genome Sequencing; 전장 유전체 분석), WES(Whole Exome Sequencing; 전장 엑솜 분석) 등이 있는데 WGS의 경우 1명의 유전체 데이터가 130GB에 달할 정도로 크기가 매우 크다. 이러한 데이터는 대부분 FASTQ 포맷으로 저장되는데 아스키코드(ASCII)로 이루어진 텍스트 기반 포맷이다.

[0005] 사람 1명의 유전체 데이터가 130GB에 이르기 때문에 데이터 저장이나 전송 부분에 있어서 압축 기술이 필수적으로 필요하다. 이에 따라 유전체 데이터를 위한 다양한 압축방식들이 제안되고 있으나, 유전체 데이터의 특성과 최근의 컴퓨팅 환경에 최적화되어 있지 않아서 압축(압축해제)에 많은 시간이 소요되는 문제가 있다.

[0006] 또한, 유전체 데이터를 전송하여 이용할 필요성이 있는데, 단순 전송하는 과정이나 압축후 전송하여 압축해제하는 과정에 많은 시간이 소요되는 문제가 있다.

[0007] 이상 종래 기술의 문제점 및 과제에 대하여 설명하였으나, 이러한 문제점 및 과제에 대한 인식은 본 발명의 기술 분야에서 통상의 지식을 가진 자에게 자명한 것은 아니다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 대한민국 공개특허 10-2002-0040406, 2002년 05월 30일 공개, "유전자 코드에 의한 정보압축 및 저장 방법"

발명의 내용

해결하려는 과제

[0009] 본 발명의 목적은 유전체 데이터의 특성과 최근의 컴퓨팅 환경에 보다 최적화된, FASTQ 포맷의 유전체 데이터 압축 방법 및 압축 전송 방법을 제공하기 위한 것이다.

[0010] 또한 본 발명의 다른 목적은 압축 및 전송에 소요되는 시간을 보다 절감할 수 있는, FASTQ 포맷의 유전체 데이터 압축 방법 및 압축 전송 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0011] 본 발명의 일 양상은, M개의 코어로 구성되는 프로세서에서 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법으로서,

[0012] M개의 코어중 하나의 코어인 제 1 코어가 첫번째 시퀀스 데이터의 제1 라인에서 고정 헤더 데이터를 압축결과 저장소에 저장하는 단계; 상기 제 1 코어가 나머지 M-1(M은 4이상의 자연수)개의 코어(이하 '기타 코어들'이라 한다) 각각에 대하여 N(N은 2이상의 자연수)개씩의 시퀀스 데이터를 분배하고 기타 코어들에서 각각 압축되도록 하여, 한번에 N*(M-1)개의 시퀀스 데이터에 대한 압축을 병렬 처리하고 상기 압축결과 저장소에 저장되도록 하는 단계;를 포함하되,

[0013] 상기한 기타 코어들의 각각에서 실행되는 압축은, 각 시퀀스 데이터에 대하여, 첫번째 라인의 고정헤더를 제거하는 과정; 두번째 라인을 부호화하는 과정; 세번째 라인의 식별자를 저장하는 과정; 네번째 라인을 반복길이 부호화하는 과정;으로 구성되는 처리 과정을 상기 N개의 시퀀스 데이터에 대하여 반복하는 제 1 단계 압축과, 상기 N개의 시퀀스 데이터에 대한 상기 제1 단계 압축의 결과에 대하여, 무손실 압축 알고리즘에 의해 압축하는 제 2 단계 압축;을 실행하는 것을 특징으로 한다.

[0014] 상기한 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법에 있어서, 상기 제 2 단계 압축에서는 7z 압축 알고리즘이 적용될 수 있다.

[0015] 상기한 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축 방법에 있어서, 상기 제 2 라인의 부호화는, 사전 설정되는 맵핑 테이블에 의해 A,T,G,C의 문자를 2bit 코드로 맵핑함으로써 실행될 수 있다.

[0016] 본 발명의 다른 양상은, M개의 코어로 구성되는 프로세서를 구비한 송신처에서 FASTQ 포맷의 유전체 데이터를 압축하여 수신처로 전송하는 유전체 데이터의 압축전송 방법으로서, 상기 송신처에서의 압축은, M개의 코어중 하나의 코어인 제 1 코어가 나머지 M-1(M은 4이상의 자연수)개의 코어(이하 '기타 코어들'이라 한다) 각각에 대

하여 N(N은 2이상의 자연수)개씩의 시퀀스 데이터를 분배하고 상기 기타 코어들에서 압축되도록 하여, 한번에 N*(M-1)개의 시퀀스 데이터에 대한 압축을 병렬 처리하는 단계;를 포함하되,

[0017] 상기한 기타 코어들의 각각에서 실행되는 압축은, 각 시퀀스 데이터에 대하여, 첫번째 라인의 고정헤더를 제거하는 과정; 두번째 라인을 부호화하는 과정; 세번째 라인의 식별자를 저장하는 과정; 네번째 라인을 반복길이 부호화하는 과정;으로 구성되는 처리 과정을 상기 N개의 시퀀스 데이터에 대하여 반복하는 제 1 단계 압축을 포함하여 실행되도록 하는 것을 특징으로 한다.

[0018] 상기한 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축전송 방법에 있어서, 상기한 기타 코어들의 각각에서 실행되는 압축은, 상기 N개의 시퀀스 데이터에 대한 상기 제 1 단계 압축의 결과에 대하여, 무손실 압축 알고리즘에 의해 압축하는 제 2 단계 압축;을 선택적으로 실행할 수 있다.

[0019] 상기한 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축전송 방법에 있어서, 상기 제 1 코어가 첫번째 시퀀스 데이터의 첫번째 라인에서 고정 헤더 데이터를 수신처로 전송하는 단계; 상기 송신처에서 유전체 데이터중 S(S는 1이상의 자연수)개의 시퀀스 데이터에 대하여 상기 제 1 단계 압축을 실행하여 상기 수신처로 전송하고 상기 수신처에서 압축해제를 실행하며, 상기 제 2 단계 압축을 실행하여 상기 수신처로 전송하고 상기 수신처에서 상기 제 2 단계 압축에 대응하는 압축해제를 실행하되, 하나 이상의 무손실 압축 알고리즘별로 실행하는 속도 측정용 압축 및 전송을 실행하는 단계; 상기 송신처의 제 1 코어가 상기 속도 측정용 압축 및 전송의 결과에 따라, 상기 S개의 시퀀스 데이터를 제외한 나머지 시퀀스 데이터(이하 '메인 시퀀스 데이터'라고 한다)에 대하여 적용할 상기 제 2 단계 압축의 실시 여부와 상기 제 2 단계 압축에 적용할 무손실 알고리즘을 결정함으로써 메인 시퀀스 데이터에 대한 압축 방식을 결정하는 단계; 상기 메인 시퀀스 데이터에 대하여 상기 결정된 압축 방식에 따라 상기 나머지 코어들에 의해 압축이 실행되도록 하며 압축된 결과가 상기 수신처로 전송되도록 할 수 있다.

[0020] 상기한 FASTQ 포맷의 유전체 데이터를 압축하는 유전체 데이터의 압축전송 방법에 있어서, 상기한 메인 시퀀스 데이터에 대한 압축 방식의 결정은, 상기 속도 측정용 압축 및 전송의 결과를 이용하여, 상기 제 1 단계 압축을 실시하고 상기 제 2 단계 압축을 실시하지 않는 압축 전송 형태와, 상기 제 1 단계 압축을 실시하고 적어도 하나 이상의 후보가 되는 무손실 압축 알고리즘에 따라 상기 제 2 단계 압축을 실시하는 압축전송 형태의 각각에 대하여, 상기 송신처 및 수신처에서 압축, 전송 및 압축해제에 소요되는 총 예상 소요시간을 계산하고 총 예상 소요시간이 가장 작은 압축전송 형태를 상기 메인 시퀀스 데이터에 대한 압축 방식으로 결정할 수 있다.

발명의 효과

[0021] 본 발명의 유전체 데이터 압축 방법에 따르면, 대용량인 유전체 데이터를 효과적으로 압축하여 스토리지에 저장하는 공간을 적게 차지하게 하고, 전송 시 고속으로 전송할 수 있게 한다. 특히 최근 NGS를 비롯한 유전체 데이터 분석 기술이 실제 활용되고 있는 바, 본 발명은 이에 대한 기반 기술로 이용될 수 있다.

[0022] 또한, 본 발명의 유전체 데이터 압축 방법에 따르면, 멀티 코어 프로세싱(병렬 컴퓨팅)의 적용이 용이하도록 된 알고리즘 구조이므로, 고압축율을 구현할 뿐만 아니라 압축에 소요되는 시간을 대폭 절약할 수 있는 효과가 있다.

[0023] 또한, 본 발명의 유전체 데이터 압축 전송 방법에 따르면, 총 예상 시간의 측정 및 계산이 용이하도록 하고 나아가 미리 예측된 총 소요시간이 나머지 시퀀스 데이터에 적용될 때의 실제 소요시간과 매우 근사한 결과를 가져올 수 있게 하며, 기존 상용의 무손실 압축 알고리즘에 대한 추가 적용 또는 비적용을 쉽게 구현할 수 있게 하며, 최적의 무손실 압축 알고리즘을 선택하는 것을 손쉽게 한다.

도면의 간단한 설명

[0024] 도 1은 FASTQ 포맷의 시퀀스 데이터 구조를 보여주는 도면이다.

도 2 및 도 3은 본 발명의 제 1 실시예에 따라 멀티코어 프로세서를 가진 서버 또는 PC에서 실행되는 유전체 데이터의 압축 방법으로서, 도 2는 어느 한 코어에서 실행되는 과정을 도시한 것이고 도 3은 나머지 코어에서 실행되는 과정을 도시한 것이다.

도 4는 본 발명의 제 2 실시예에 따라 M개의 코어로 구성되는 프로세서를 구비한 송신처에서 FASTQ 포맷의 유전체 데이터를 압축하여 수신처로 전송하는 유전체 데이터의 압축전송 방법을 도시한 플로우차트이다.

도 5는 병렬 처리없이, 예를 들면 컴퓨터의 단일 코어에서 유전체 데이터를 압축하는 실험예에 따른 실험결과

데이터를 보여준다.

발명을 실시하기 위한 구체적인 내용

[0025] 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 명칭 및 도면 부호를 사용한다.

[0027] 도 1은 FASTQ 포맷의 시퀀스 데이터 구조를 보여주는 도면이다.

[0028] 각 시퀀스 데이터에서 첫번째 라인(Line 1)은 고정 헤더 데이터와 변경 헤더 데이터를 포함하며, 예를 들어 첫번째 라인은 "@HWUSI-EAS100R:6:73:941:1973#0/1"와 같이 표현될 수 있다.

[0029] 'HWUSI-EAS100R'는 the unique instrument name을 표시하고, '6'은 flowcell lane, '73'은 tile number within the flowcell lane, '941'는 'x'-coordinate of the cluster within the tile, '1973'은 'y'-coordinate of the cluster within the tile, '#0'는 index number for a multiplexed sample (0 for no indexing), '/1'은 the member of a pair(/1 or /2; paired-end or mate-pair reads only)을 표시한다.

[0030] 시퀀스 데이터의 두번째 라인(Line 2)은 실제 시퀀스 데이터(원시 시퀀스 데이터)로서 A, T, G, C 4가지 종류가 반복되는 텍스트 데이터이다. 예를 들면 두번째 라인은 "GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT"와 같이 표현될 수 있다.

[0031] 세번째 라인(Line 3)은 1byte의 식별자로서 '+'이며, 네 번째 라인(Line 4)은 시퀀스 데이터의 QV(quality value) 정로서, 예를 들면, "!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65"와 같이 표현된다.

[0032] 본 발명은 FASTQ 포맷의 유전체 데이터 파일을 압축하여 저장하거나 압축하여 전송하는 방법에 관한 것으로서, 이러한 압축 방법은 서버, 워크스테이션 또는 PC 등에서 실행될 수 있으며, 나아가 멀티 코어를 가진 마이크로프로세서를 구비한 서버, 워크스테이션 또는 PC 등에서 실행될 수 있다. 또한, 압축된 유전체 데이터 파일은 인터넷과 같은 통신망을 통하여 이들 사이에서 전송될 수 있다.

[0034] 도 2 및 도 3은 본 발명의 제 1 실시예에 따라 멀티코어 프로세서를 가진 서버 또는 PC에서 실행되는 유전체 데이터의 압축 방법으로서, 도 2는 어느 한 코어에서 실행되는 과정을 도시한 것이고 도 3은 나머지 코어에서 실행되는 과정을 도시한 것이다.

[0035] 예를 들어, M개의 코어로 구성되는 프로세서로써, FASTQ 포맷의 유전체 데이터를 압축하기 위하여, M개의 코어 중 하나의 코어(이하, '제 1 코어'라고도 한다)에서는 도 2에 도시된 과정을 실행하고 제 1 코어를 제외한 나머지 M-1(M은 4이상의 자연수)개의 코어(이하 '기타 코어들'이라고도 한다)는 도 3에 도시된 과정을 실행하며, 이하에서는 본 발명의 실시예를 설명하기 위한 사항에 집중하기 위해 기타 부수적인 사항의 설명은 생략될 수 있다.

[0036] 먼저, 제 1 코어는 수많은 시퀀스 데이터에서 첫번째 시퀀스 데이터의 제 1 라인(Line 1)에서 고정 헤더 데이터를 압축결과 저장소에 저장한다(S11). 그리고 제 1 코어는 기타 코어들에 대하여 각 코어마다 N개(N은 2이상의 자연수)의 시퀀스 데이터를 분배하는 데(S12), 이와 유사하게 각 코어가 순서대로 N개씩의 시퀀스 데이터를 리드하는 것도 이와 균등한 것으로 본다. 이와 같이 기타 코어들에게 분배(분할)된 N개의 시퀀스 데이터는 도 3에 도시된 바와 같은 압축과정을 실행한다. 제 1 코어는 나머지 M-1개의 코어 각각에 대하여 N개씩의 시퀀스 데이터를 분배하고 기타 코어들에서 각각 압축되도록 한다.

[0037] 이에 따라 프로세서는 한번에 N*(M-1)개의 시퀀스 데이터에 대한 압축을 병렬 처리하며, 기타 코어들에서 압축된 결과는 압축결과 저장소에 저장되도록 한다(S13). 제 1 코어 또는 각 기타 코어들은 압축된 결과를 압축결과 저장소에 저장되도록 하며, 결국 압축된 N*(M-1)개의 시퀀스 데이터가 저장된다(S13).

[0038] 도 3을 참조하면서, 기타 코어들의 각각에서 실행되는 압축 과정을 구체적으로 살펴보면, 각 시퀀스 데이터에 대하여, 첫번째 라인(Line 1)의 데이터에 있어서 고정 헤더 데이터를 제거하며(S21, S31), 변동 헤더 데이터는 그대로 남겨둔다. 추후 압축해제시에는 제 1 코어가 저장한 첫번째 시퀀스 데이터의 고정 헤더 데이터를 이용하여 각 시퀀스 데이터의 고정 헤더 데이터를 복구한다.

[0039] 두번째 라인(Line 2)의 원시 시퀀스 데이터(또는 원시 시퀀스 문자열(Raw Sequence Letters)라고도 한다)에 대해서는 부호화(S22,S32)를 수행하는 데, 사전 설정되는 맵핑 테이블에 의해 A,T,G,C의 문자를 2bit 코드로 맵핑함으로써 실행된다. 따라서 1 Byte(8 bit) 데이터는 2 bit 데이터가 된다. 맵핑 테이블은 A,T,G,C의 문자에 대하여 대응하는 2bit 코드를 정의하는 테이블이되며, 추후 압축해제시에는 맵핑 테이블을 이용하여 2 bit 코드에 대응하는 문자로 복구한다. 세번째 라인의 식별자는 그대로 남겨서 저장하며(S23,S33), 네번째 라인의 QV 데이터는 반복길이 부호화(Run Length Encoding) 기법을 적용해서(S24,S34) 압축한다.

[0040] 그리고 이와 같이 하나의 시퀀스 데이터에 대한 과정을 반복하여 N개의 시퀀스 데이터에 대하여 압축하는 데, 예를 들어, N개의 시퀀스에 대한 처리가 완료되었는지를 판단하여(S25,S35), 아니면 다음 첫번째 라인 부터 네번째 라인까지의 처리를 수행하고 처리가 완료되었으면 제 1 단계 압축을 마치게 되고 제 2 단계 압축을 실행하게 된다.

[0041] N개의 시퀀스 데이터에 대한 제1 단계 압축의 결과 데이터에 대하여, 무손실 압축 알고리즘을 적용함으로써 제 2 단계 압축을 실행한다(S26). 제 2 단계 압축에서는 7z(7zip) 압축 알고리즘이 적용된다.

[0042] 상기의 과정들은 M-1개의 코어에서 각각 실행되며, 각 코어들로부터의 압축 결과 데이터는 압축 결과 저장소에 저장된다. 이와 같은 방식으로 기타 코어들로의 분배 및 압축을 반복하며, 제 1 코어는 전체 유전체 데이터 파일에 대한 압축이 완료되었는지를 판단하고(S14), 완료되지 않았으면 단계 S12 내지 단계 S14를 반복하여 유전체 데이터 파일의 모든 시퀀스 데이터에 대하여 실행되도록 한다.

[0044] 도 5는 병렬 처리없이, 예를 들면 컴퓨터의 단일 코어에서 유전체 데이터를 압축하는 실험예에 따른 실험결과 데이터를 보여준다.

[0045] 도 5에서 첫번째 컬럼의 압축방법 cfc는 각 시퀀스 데이터에 대하여 상기한 제 1 단계 압축(단계 S21 내지 단계 S24)만을 반복수행하여 유전체 데이터의 전체를 압축한 예이고, 압축방법 7z는 동일 예의 유전체 데이터에 대해 7z 압축 알고리즘만을 적용하여 압축한 예이며, 압축방법 7z[cfc]는 cfc 압축방법을 통해 얻은 결과가 있는 것을 전제로 이러한 결과에 대하여 제 2 단계 압축(7z 압축 알고리즘)을 수행한 예이고, 압축방법 cfc+7z는 제 1 단계 압축과 제 2 단계 압축을 모두 수행한 결과를 표시한 예이다.

[0046] 도시된 바와 같이, 7z만을 이용하여 압축하는 경우는 cfc 압축방법(제 1 단계 압축 방법만을 이용하는 방식)에 비하여 압축률은 좋지만, 압축시간이 약 7시간으로서 cfc의 35분에 비하여 소요시간이 많이 걸린다. 바꾸어 말하면 cfc 압축방법은 압축시간이 현존하는 최고수준의 압축률과 속도를 가진 7z에 비해서도 압축시간이 대폭 절약되어서 절대적 우위를 가지나, 압축후 크기에 있어서 대략 3배에 육박하는 크기를 가져서 저장 공간이나 전송량의 관점에서 약점을 가진다.

[0047] 한편, 본 발명의 실시예에 적용된 제 1 단계 압축 및 제 2 단계 압축을 결합한 cfc+7z 압축방법은 cfc 압축방법에 비해서는 시간이 많이 걸리나, 7z 압축방법에 비해서는 압축 시간을 절약하는데, 특히, 7z 압축방법에 비해서 압축 시간과 압축률의 모두에서 더 좋은 장점이 있다.

[0048] 나아가, 단일 코어를 이용한 상기한 실험예의 cfc+7z 압축방법은 압축시간에 있어서 대폭적인 단축이 어려웠던 과제가 그대로 남아 있으나, 상기한 도 2 및 도 3과 같이 본 발명의 일 실시예에 따른 압축방법은 멀티 코어 프로세싱의 적용이 용이하도록 설계된 알고리즘이므로, 멀티 코어 프로세싱을 수행하면 실행 시간이 대략 1/(M-1)에 근접할 정도로 줄어들며, 예를 들어 8 코어 프로세서를 이용하면 대략 1/7의 압축 시간으로 절약될 수 있다. 기존 유전체 데이터의 압축방법에 비하여 본 발명의 일 실시예에 따른 유전체 데이터의 압축방법은 멀티코어 프레세싱에 매우 용이하게 적용될 수 있는 장점이 있다.

[0050] 도 4는 본 발명의 제 2 실시예에 따라 M개의 코어로 구성되는 프로세서를 구비한 송신처에서 FASTQ 포맷의 유전체 데이터를 압축하여 수신처로 전송하는 유전체 데이터의 압축전송 방법을 도시한 플로우차트이다.

[0051] 송신처는 유전체 데이터를 압축하여 수신처로 전송하며 수신처에서는 유전체 데이터를 수신하여 압축해제한다. 송신처 및 수신처는 서버, 워크스테이션 또는 PC일 수 있다.

[0052] 송신처는, 예를들어 송신처에 있는 프로세서(그 중 제 1 코어)의 제어로, 유전체 데이터에서 첫번째 시퀀스 데이터의 첫번째 라인에서 고정 헤더 데이터를 수신처로 전송하도록 하며(S41), 이러한 과정은 단계 S42 내지 S45 중 어느 한 과정의 전후로 이동되어도 된다.

[0053] 그리고, 송신처에서는 유전체 데이터중 S(S는 1이상의 자연수)개의 시퀀스 데이터에 대하여 상기한 제 1 단계 압축을 실행하여 수신처로 전송하고 수신처에서 압축해제를 실행하며, 제 1 단계 압축 결과 데이터에 대해 상기

한 제 2 단계 압축을 실행하여 수신처로 전송하고 수신처에서 제 2 단계 압축에 대응하는 압축해제를 실행하되, 하나 이상, 복수의 무손실 압축 알고리즘별로 실행함으로써, 속도 측정용 압축 및 전송과 소요시간 측정을 실행한다(S42). 즉, 제 1 단계 압축 만을 실행한 결과, 무손실 압축 알고리즘별로 제 2 단계 압축까지 실행한 결과를 전송하여 각각에 대하여 걸리는 전체 시간을 측정한다. 예를 들어, 송신처는 자체적으로 압축 및 전송에 걸리는 시간을 측정할 수 있으며, 수신처가 측정한 압축해제 시간을 보고받을 수 있다. 예를 들어, 복수의 무손실 압축 알고리즘은 ZIP, RAR, 7z, GZ, BZ2, ALZ, EGG, Raw, LHA, ARJ, ACE 등의 전부 또는 일부를 포함하는 집합일 수 있다.

[0054] 상기한 속도 측정용 압축 및 전송의 결과에 따라, 측정용을 겸했던 S개의 시퀀스 데이터를 제외한 나머지 시퀀스 데이터(이하 '메인 시퀀스 데이터'라고 한다)에 대하여 적용할 제 2 단계 압축의 실시 여부와 제 2 단계 압축에 적용할 무손실 알고리즘을 결정함으로써 메인 시퀀스 데이터에 대한 압축 방식을 결정한다.

[0055] 송신처, 예를 들면 송신처의 제 1 코어는, 상기한 속도 측정용 압축 및 전송의 결과를 이용하여, 도 3을 통해 설명되었던 제 1 단계 압축을 실시하고 상기한 제 2 단계 압축을 실시하지 않는 압축 전송 형태와, 제 1 단계 압축을 실시하고 적어도 하나 이상의 후보가 되는 무손실 압축 알고리즘에 따라 제 2 단계 압축을 실시하는 압축전송 형태의 각각에 대하여, 송신처 및 수신처에서 압축, 전송 및 압축해제에 소요되는 총 예상 소요시간을 계산하고(S43) 총 예상 소요시간이 가장 작은 압축전송 형태를 메인 시퀀스 데이터에 대한 압축 방식으로 결정한다(S44).

[0056] 그리고, 메인 시퀀스 데이터에 대하여 결정된 압축 방식에 따라 나머지 코어들에 의해 도 3에서 도시된 바와 같이 압축이 실행되도록 하며 압축된 결과가 수신처로 전송되도록 한다.

[0057] 송신처에서의 압축은, M개의 코어중 하나의 코어인 제 1 코어가 나머지 M-1(M은 4이상의 자연수)개의 코어(이하 '기타 코어들'이라 한다) 각각에 대하여 N(N은 2이상의 자연수)개씩의 시퀀스 데이터를 분배하고 기타 코어들에서 압축되도록 하여, 한번에 N*(M-1)개의 시퀀스 데이터에 대한 압축을 병렬 처리하는 단계를 수행한다. 이때 기타 코어들의 각각에서 실행되는 압축은, 도 3의 설명에서와 같이, 각 시퀀스 데이터에 대하여, 첫번째 라인의 고정헤더를 제거하고(S21,S31), 변동헤더를 저장하거나 그대로 두는 과정, 두번째 라인을 부호화하는 과정(S22, S23), 세번째 라인의 식별자를 그대로 두거나 저장하는 과정(S23, S33), 네번째 라인을 반복길이 부호화하는 과정(S24, S34)을 포함하여 구성되는 처리 과정을 가지며, 이러한 과정을 N개의 시퀀스 데이터에 대하여 반복하는 제 1 단계 압축을 포함하여 실행된다.

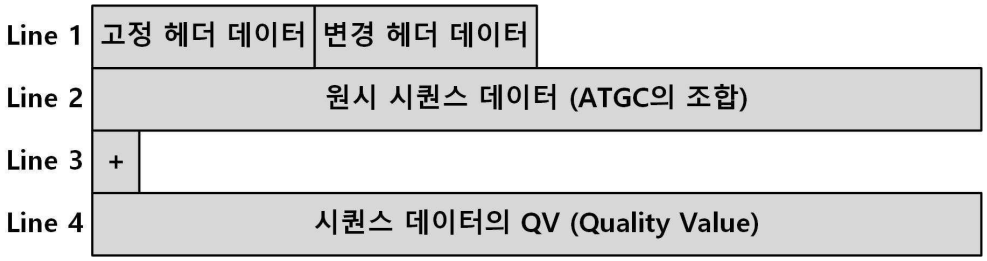
[0058] 그리고, 상기 N개의 시퀀스 데이터에 대한 상기 제 1 단계 압축의 결과에 대하여, 결정된 무손실 압축 알고리즘에 의해 압축하는 제 2 단계 압축을 선택적으로 실행한다.

[0059] 본 발명의 제 2 실시예에 따르면 제 2 단계 압축은 상황에 따라 선택적으로(다이내믹하게) 실행된다. 예를 들어, 통신을 통한 전송 속도가 낮은 환경이나 상황에서는 제 2 단계 압축까지 수행되는 것으로 하여 단계 S44에서 압축 방식 결정이 이루어질 수 있다. 전송 속도가 낮은 경우에는 압축을 보다 많이 하여 전송함으로써 전체 소요 시간을 단축할 수 있다. 그러나, 전송 속도가 일정 수준 이상이 되면 제 2 단계 압축 및 압축 해제에 걸리는 시간 증가가 이러한 전송 시간 단축보다 더 길게 될 것이므로 제 1 단계 압축만을 실행하는 것으로 결정될 수도 있다. 본 발명의 실시예에 따르면 이러한 결정은 자동으로 수행된다.

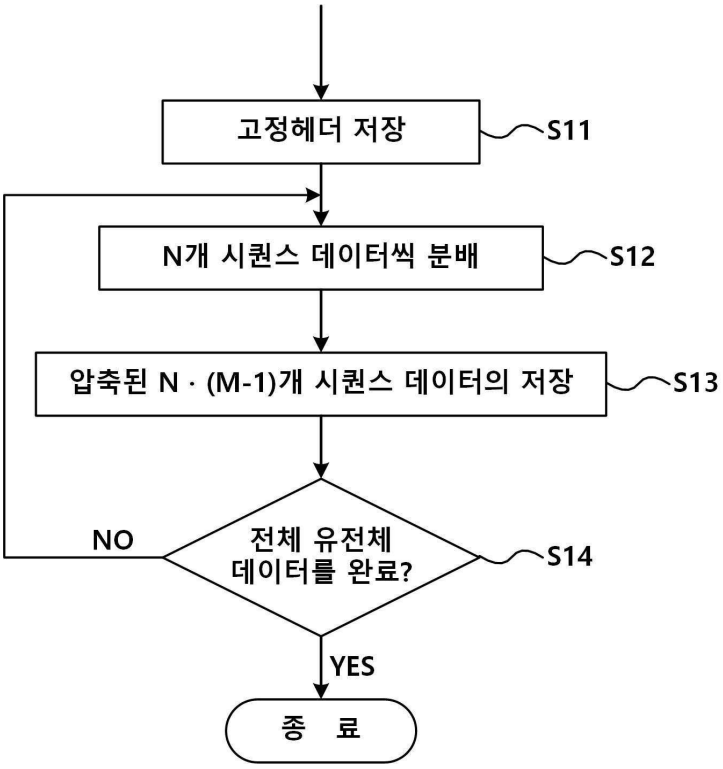
[0060] 본 발명에 따른 유전체 데이터의 압축 전송 방법은 이와 같은 총 예상 시간의 측정 및 계산이 용이하도록 하고 나아가 미리 예측된 총 소요시간이 나머지 시퀀스 데이터에 적용될 때의 실제 총 소요시간과 매우 근사한 결과를 가져올 수 있게 한다. 나아가, 본 발명에 따른 유전체 데이터의 압축 전송 방법은 기존 상용의 무손실 압축 알고리즘에 대한 추가 적용 또는 비적용을 쉽게 구현할 수 있게 하며, 최적의 무손실 압축 알고리즘을 선택하는 것을 손쉽게 한다.

도면

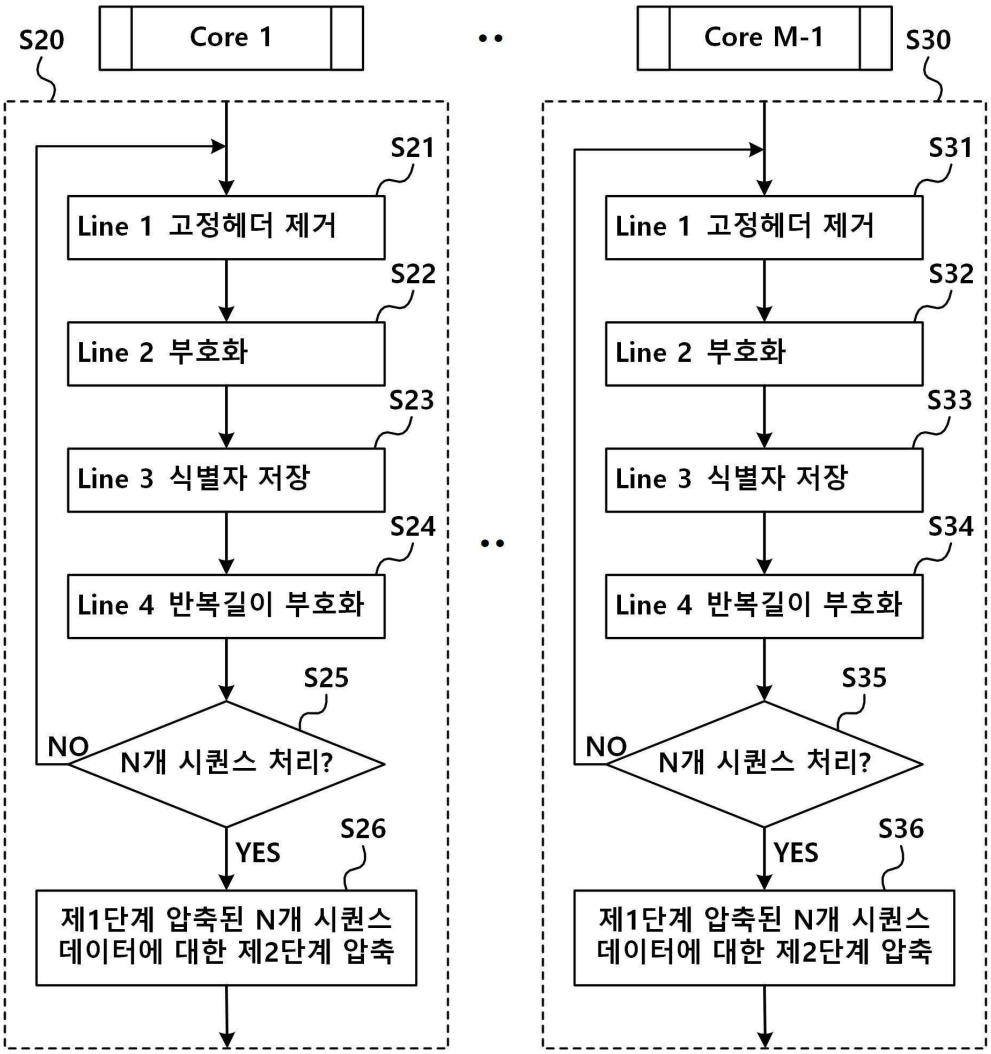
도면1


Line 1
고정 헤더 데이터
변경 헤더 데이터
Line 2
원시 시퀀스 데이터 (ATGC의 조합)
Line 3
+
Line 4
시퀀스 데이터의 QV (Quality Value)

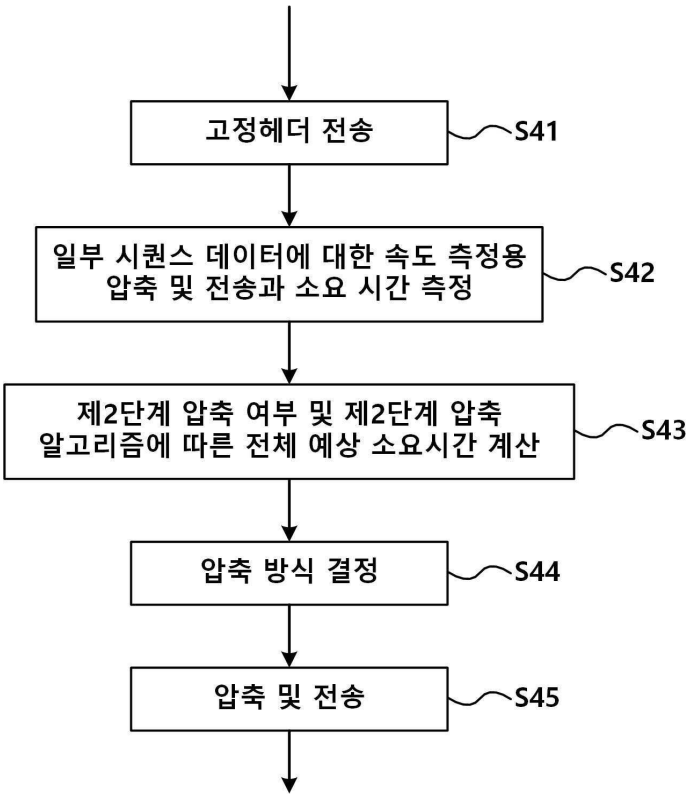
도면2


고정헤더 저장
S11
N개 시퀀스 데이터씩 분배
S12
압축된 N · (M-1)개 시퀀스 데이터의 저장
S13
NO
전체 유전체
데이터를 완료?
S14
YES
종 료

도면3


Core 1
Core M-1
S20
S30
S21
S31
Line 1 고정헤더 제거
S22
S32
Line 2 부호화
S23
S33
Line 3 식별자 저장
S24
S34
Line 4 반복길이 부호화
S25
S35
NO
N개 시퀀스 처리?
YES
S26
S36
제1단계 압축된 N개 시퀀스 데이터에 대한 제2단계 압축

도면4



도면5

압축방법	압축 전 크기	압축 후 크기	압축률	시간
cfc	144G	75G	47.92%	약 35분
7z	144G	28G	80.56%	약 7시간
7z [cfc]	75G	25.9G	65.46%	약 5시간
cfc+7z	144G	25.9G	82.01%	약 5시간 30분