



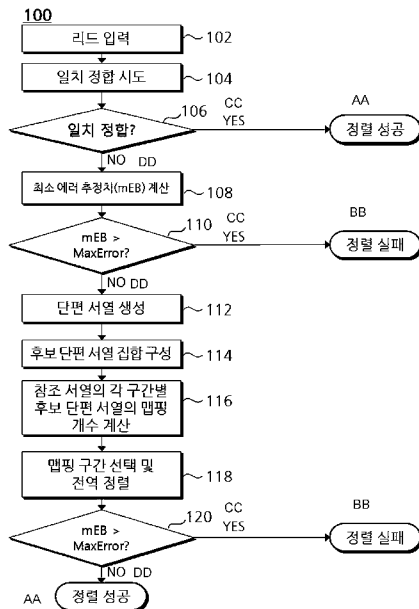
- (51) 국제특허분류: *C12Q 1/68* (2006.01) *G06F 19/22* (2011.01)
- (21) 국제출원번호: PCT/KR2013/007276
- (22) 국제출원일: 2013년 8월 13일 (13.08.2013)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보: 10-2012-0120448 2012년 10월 29일 (29.10.2012) KR
- (71) 출원인: 삼성에스디에스 주식회사 (SAMSUNG SDS CO., LTD.) [KR/KR]; 135-918 서울시 강남구 역삼2동 707-19 일옥빌딩, Seoul (KR).
- (72) 발명자: 박민서 (PARK, Minseo); 135-284 서울시 강남구 대치4동 931-6, Seoul (KR).
- (74) 대리인: 노준태 (NOH, Jun Tae) 등; 137-953 서울시 서초구 서초중앙로 53 6층 (서초동, 대림빌딩), Seoul (KR).

- (81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

[다음 쪽 계속]

(54) Title: BASE SEQUENCE ALIGNMENT SYSTEM AND METHOD

(54) 발명의 명칭: 염기 서열 정렬 시스템 및 방법



102 ... Input read
104 ... Attempt exact matching
106 ... Exact match?
108 ... Calculate minimum error estimated value (mEB)
112 ... Generate fragment sequence
114 ... Constitute candidate fragment sequence collection
116 ... Calculate candidate fragment sequence mapping number, separately for section in reference sequence
118 ... Select and globally align mapping section
AA ... Alignment successful
BB ... Alignment failed
CC ... YES
DD ... NO

(57) Abstract: Disclosed are a base sequence alignment system and method. According to one embodiment of the present invention, the base sequence alignment system comprises: a fragment sequence generating unit which generates a plurality of fragment sequences from a read; a filtering unit which constitutes a candidate fragment sequence collection comprising only fragment sequences matching a reference sequence from among the plurality of fragment sequences generated; a mapping number calculating unit which divides the reference sequence into a plurality of sections, and calculates the total mapping number of the candidate fragment sequences, separately for each section; and an alignment unit which selects a section in which the calculated total mapping number is equal to or greater than a reference number, and implements global alignment of the read with respect to the selected section.

(57) 요약서: 염기 서열 정렬 시스템 및 방법이 개시된다. 본 발명의 일 실시예에 따른 염기 서열 정렬 시스템은, 리드로부터 복수 개의 단편(fragment) 서열들을 생성하는 단편 서열 생성부, 생성된 상기 복수 개의 단편 서열들 중 참조 서열과 매칭되는 단편 서열만을 포함하는 후보 단편 서열 집합을 구성하는 필터링부, 상기 참조 서열을 복수 개의 구간으로 분할하고, 상기 각 구간 별 상기 후보 단편 서열들의 총 매핑 개수를 계산하는 매핑 개수 계산부, 및 계산된 상기 총 매핑 개수가 기준 개수 이상인 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역 정렬(global alignment)을 수행하는 정렬부를 포함한다.

공개:

— 국제조사보고서와 함께 (조약 제 21 조(3))

명세서

발명의 명칭: 염기 서열 정렬 시스템 및 방법

기술분야

- [1] 본 발명의 실시예들은 유전체의 염기 서열을 분석하기 위한 기술과 관련된다.

배경기술

- [2] 저렴한 비용과 빠른 데이터 생산으로 인해 대용량의 짧은 서열을 생산하는 차세대 시퀀싱(NGS; Next Generation Sequencing)이 전통적인 생거(Sanger) 시퀀싱 방식을 빠르게 대체하고 있다. 또한 다양한 NGS 서열재조합 프로그램들이 정확도에 초점을 맞추어 개발되었다. 그러나, 최근 차세대 시퀀싱 기술이 발전함에 따라 단편 서열을 만들어 내는 비용이 예전의 절반 이하가 되었고, 이에 따라 사용할 수 있는 데이터의 양이 많아지게 되어서, 대용량의 짧은 서열들을 빠른 시간에 정확하게 처리하기 위한 기술이 필요하게 되었다.
- [3] 서열 재조합의 첫 번째 단계는 염기 서열 정렬(alignment) 알고리즘을 통해 리드를 참조 서열의 정확한 위치에 맵핑(mapping)하는 것이다. 여기서의 문제점은 같은 종의 개체라 할지라도 다양한 유전적 변이로 인해 유전체 서열에 차이가 있을 수 있다는 점이다. 또한 시퀀싱 과정에서의 오류로 인해서도 염기 서열에 차이가 생길 수 있다. 따라서 염기 서열 정렬 알고리즘은 이러한 차이와 변이를 효과적으로 고려해서 맵핑 정확도를 높이지 않으면 안 된다.
- [4] 결론적으로, 유전체 정보의 분석을 진행하기 위해서는, 될 수 있는 한 많은 수의 정확한 전체 유전체 정보 데이터가 필요하다. 또 이를 위해서는 무엇보다도 뛰어난 정확도와 큰 처리량을 갖는 염기 서열 정렬 알고리즘을 개발하는 것이 선행되어야 한다. 그러나 종래의 방법들은 이러한 요구 조건들을 만족시키는 데 한계가 있었다.

발명의 상세한 설명

기술적 과제

- [5] 본 발명의 실시예들은 맵핑 정확도를 보장하는 동시에 맵핑 시의 복잡도를 개선하여 처리 속도를 높일 수 있는 염기 서열 정렬 수단을 제공하는 데 그 목적이 있다.

과제 해결 수단

- [6] 상기 과제를 해결하기 위한 본 발명의 일 실시예에 따른 염기 서열 정렬 시스템은, 리드로부터 복수 개의 단편(fragment) 서열들을 생성하는 단편 서열 생성부, 생성된 상기 복수 개의 단편 서열들 중 참조 서열과 매칭되는 단편 서열만을 포함하는 후보 단편 서열 집합을 구성하는 필터링부, 상기 참조 서열을 복수 개의 구간으로 분할하고, 상기 각 구간 별 상기 후보 단편 서열들의 총 맵핑 개수를 계산하는 맵핑 개수 계산부, 및 계산된 상기 총 맵핑 개수가 기준 개수 이상인 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역

정렬(global alignment)을 수행하는 정렬부를 포함한다.

- [7] 한편, 상기 과제를 해결하기 위한 본 발명의 일 실시예에 따른 염기 서열 정렬 방법은, 단편 서열 생성부에서, 리드로부터 복수 개의 단편(fragment) 서열들을 생성하는 단계, 필터링부에서, 생성된 상기 복수 개의 단편 서열들 중 참조 서열과 매칭되는 단편 서열만을 포함하는 후보 단편 서열 집합을 구성하는 단계, 맵핑 개수 계산부에서, 상기 참조 서열을 복수 개의 구간으로 분할하고, 상기 각 구간 별로 상기 후보 단편 서열들의 총 맵핑 개수를 계산하는 단계, 및 정렬부에서, 계산된 상기 총 맵핑 개수가 기준 개수 이상인 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역 정렬(global alignment)을 수행하는 단계를 포함한다.

발명의 효과

- [8] 본 발명의 실시예들에 따른 경우 리드의 정렬 시 리드의 특정 영역만을 고려하는 것이 아니라 리드 전체를 고려하여 시드(단편 서열)를 선택하므로 리드의 일부분만을 고려하는 알고리즘에 비해 정확도를 향상할 수 있다.
- [9] 또한 단편 서열 별로 참조 서열 내에서의 반복수를 제한하고, 이를 초과하는 시드에 대해서는 시드의 길이를 확장함으로써 맵핑의 정확도를 높임과 동시에 속도 또한 향상할 수 있는 효과가 있다.
- [10] 또한 참조 서열을 복수 개의 영역으로 나눈 뒤, 이 중 리드가 맵핑될 가능성이 높은 특정 영역을 선택하고, 해당 영역 내에서만 전역 정렬(Global Alignment)을 수행함으로써 전역 정렬 시간을 대폭 감소시킬 수 있다.
- [11] 아울러 리드에서 도출된 단편 서열들의 맵핑 위치와 조합을 찾는 복잡한 과정 대신, 조합을 구성할 가능성이 높은 단편 서열에 대해 바로 전역 정렬을 수행함으로써 전역 정렬 속도를 더욱 높일 수 있으며, 전역 정렬 위치를 기억하여 두었다가 해당 위치 주변에서는 반복하여 전역 정렬을 수행하지 않도록 구성함으로써 불필요한 전역 정렬 횟수를 감소시킬 수 있는 장점이 있다.

도면의 간단한 설명

- [12] 도 1은 본 발명의 일 실시예에 따른 염기 서열 정렬 방법(100)을 설명하기 위한 도면이다.
- [13] 도 2는 본 발명의 일 실시예에 따른 염기 서열 정렬 방법(100)의 108 단계에서의 최소 에러 추정치 계산 과정을 예시하기 위한 도면이다.
- [14] 도 3은 본 발명의 일 실시예에 따른 염기 서열 정렬 방법(100)의 112 단계에서의 단편 서열 생성 과정을 설명하기 위한 도면이다.
- [15] 도 4는 본 발명의 일 실시예에 따른 참조 서열 내에서의 맵핑 대상 구간 선택 과정을 예시한 도면이다.
- [16] 도 5는 본 발명의 일 실시예에 따른 전역 정렬 시 불필요한 전역 정렬 횟수를 감소시키기 위한 방법을 설명하기 위한 예시도이다.
- [17] 도 6은 본 발명의 일 실시예에 따른 염기 서열 정렬 시스템(600)을 도시한

블록도이다.

발명의 실시를 위한 형태

- [18] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.
- [19] 본 발명을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.
- [20] 본 발명의 기술적 사상은 청구범위에 의해 결정되며, 이하의 실시예는 본 발명의 기술적 사상을 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 효율적으로 설명하기 위한 일 수단일 뿐이다.
- [21]
- [22] 본 발명의 실시예들을 상세히 설명하기 앞서, 먼저 본 발명에서 사용되는 용어들에 대하여 설명하면 다음과 같다.
- [23] 먼저, "리드(read) 서열"(또는 줄여서 "리드"로 지칭)이란 게놈 시퀀서(genome sequencer)에서 출력되는 짧은 길이의 염기 서열 데이터이다. 리드의 길이는 게놈 시퀀서의 종류에 따라 일반적으로 35~500bp(base pair) 정도로 다양하게 구성되며, 일반적으로 DNA 염기의 경우 A, C, G, T의 알파벳 문자로 표현된다.
- [24] "참조 서열(reference sequence)"이란 상기 리드들로부터 전체 염기 서열을 생성하는 데 참조가 되는 염기 서열을 의미한다. 염기 서열 분석에서는 게놈 시퀀서에서 출력되는 다량의 리드들을 참조 서열을 참조하여 맵핑함으로써 전체 염기 서열을 완성하게 된다. 본 발명에서 상기 참조 서열은 염기 서열 분석 시 미리 설정된 서열(예를 들어 인간의 전체 염기 서열 등)일 수도 있으며, 또는 게놈 시퀀서에서 만들어진 염기 서열을 참조 서열로 사용할 수도 있다.
- [25] "베이스(base)"는 참조 서열 및 리드를 구성하는 최소 단위이다. 전술한 바와 같이 DNA 염기의 경우 A, C, G 및 T의 네 종류의 알파벳 문자로 구성될 수 있으며, 이들 각각을 베이스라 표현한다. 다시 말해 DNA 염기의 경우 4개의 베이스로 표현되며, 이는 리드 또한 마찬가지이다.
- [26] "단편(fragment) 서열"(또는 시드(seed))이란 리드의 맵핑을 위하여 리드와 참조 서열을 비교할 때의 단위가 되는 시퀀스이다. 이론적으로 리드를 참조 서열에 맵핑하기 위해서는 리드 전체를 참조 서열의 가장 첫 부분부터 순차적으로 비교해 나가면서 리드의 맵핑 위치를 계산하여야 한다. 그러나 이와 같은 방법의 경우 하나의 리드를 맵핑하는 데 너무 많은 시간 및 컴퓨팅 파워가 요구되므로, 실제로는 리드의 일부분으로 구성된 조각인 단편 서열을 먼저 참조 서열에 맵핑함으로써 전체 리드의 맵핑 후보 위치를 찾아 내고 해당 후보 위치에 전체

리드를 맵핑(Global Alignment)하게 된다.

[27]

[28] 도 1은 본 발명의 일 실시예에 따른 염기 서열 정렬 방법(100)을 설명하기 위한 도면이다. 본 발명의 실시예에서, 염기 서열 정렬 방법(100)이란 게놈 시퀀서(genome)에서 출력되는 리드를 참조 서열과 비교하여 리드의 상기 참조 서열에서의 맵핑(또는 정렬) 위치를 결정하는 일련의 과정을 의미한다.

[29] 먼저, 게놈 시퀀서(genome sequencer)로부터 리드가 입력되면(102), 리드 전체와 상기 참조 서열과의 일치 정합(exact matching)을 시도한다(104). 만약 상기 시도 결과 리드 전체에 대한 일치 정합이 성공한 경우에는 이후의 정렬 단계를 수행하지 않고 정렬에 성공한 것으로 판단한다(106). 인간의 염기 서열을 대상으로 한 실험 결과, 게놈 시퀀서에서 출력되는 100만 개의 리드를 인간의 염기 서열에 일치 정합할 경우 총 200만회의 정렬 중(정방향 시퀀스 100만회, 역상보(reverse complement) 방향 시퀀스 100만회) 231,564회의 일치 정합이 발생하는 것으로 나타났다. 따라서 상기 104 단계의 수행 결과 약 11.6%만큼의 정렬 소요량을 감소시킬 수 있었다.

[30] 그러나, 이와 달리 상기 106 단계에서 해당 리드가 일치 정합되지 않는 것으로 판단되는 경우에는 해당 리드를 상기 참조 서열에 정렬했을 때 나타날 수 있는 최소 에러 추정치(mEB; minimum Error Bound)를 계산한다(108).

[31] 도 2는 상기 108 단계에서의 최소 에러 추정치 계산 과정을 예시하기 위한 도면이다. 먼저, 도시된 바와 같이 최초 최소 에러 추정치를 0으로 설정하고($mEB = 0$) 리드의 가장 첫 번째 베이스부터 오른쪽으로 한 베이스씩 이동하면서 일치 정합을 시도한다. 이때 상기 리드의 특정 베이스(도면에서 왼쪽에서 첫번째 화살표)에서부터 더 이상 매칭이 불가능하다고 가정하자. 이 경우는 리드의 정합 시작 위치부터 현재 위치 사이의 구간 어딘가에서 에러가 발생한 것을 의미한다. 따라서 이 경우에는 최소 에러 추정치를 1만큼 증가시키고($mEB = 1$), 다음 위치에서 새로 일치 정합을 시작한다. 이후 또 다시 일치 정합이 불가능하다고 판단되는 경우에는, 일치 정합을 새로 시작한 위치부터 현재 위치 사이의 구간 어디에서 다시 에러가 발생한 것이므로, 최소 에러 추정치를 다시 1만큼 증가시키고($mEB = 2$), 다음 위치에서 새로 일치 정합을 시작한다. 이와 같은 과정을 거쳐 리드의 끝까지 도달한 경우의 최소 에러 추정치(도면에서 $mEB = 3$)가 해당 리드에서 발생할 수 있는 에러의 개수가 된다. 이때 상기 mEB 값이 최소 에러 추정치인 이유는 리드에서 에러가 발생할 수 있는 모든 에러의 수를 검토한 것이 아니라, 일정 부분에서 에러가 발생하면 그 이후부터 새로 일치 정합을 수행하는 방식으로, 대상 서열의 한 위치(position)에 대해서만 검사하기 때문이다. 즉, 상기 mEB 값은 해당 리드에서 발생할 수 있는 에러의 최소값이 될 수 있으며, 대상 서열의 다른 위치에서는 더 많은 에러가 나올 수 있다.

[32] 상기과 같은 과정을 거쳐 리드의 최소 에러 추정치가 계산되면, 계산된 최소 에러 추정치가 기 설정된 최대 에러 허용치(MaxError)를 초과하는지의 여부를

판단하고(110), 초과하는 경우 해당 리드에 대한 정렬이 실패한 것으로 판단하여 정렬을 종료한다. 전술한 인간의 염기 서열을 대상으로 한 실험에서, 최대 에러 허용치(MaxError)를 3으로 하고 나머지 리드들의 최소 에러 추정치를 계산한 결과, 총 844,891회에 해당하는 리드들이 상기 최대 에러 허용치를 초과하는 것으로 나타났다. 즉, 상기 108 단계의 수행 결과 약 42.2%만큼의 정렬 소요량을 감소시킬 수 있었다.

[33] 그러나 이와 달리 상기 110 단계에서의 판단 결과, 계산된 최소 에러 추정치가 상기 최대 에러 허용치 이하인 경우에는 다음과 같은 과정을 거쳐 해당 리드에 대한 정렬을 수행한다.

[34] 먼저, 상기 리드로부터 복수 개의 단편(fragment) 서열들을 생성하고(112), 생성된 상기 복수 개의 단편 서열들 중 상기 참조 서열과 매칭되는 단편 서열만을 포함하는 후보 단편 서열 집합을 구성한다(114). 이후 상기 참조 서열을 복수 개의 구간으로 분할하고, 상기 각 구간 별로 상기 후보 단편 서열들의 총 맵핑 개수를 계산하며(116), 상기 계산 결과에 따라 총 맵핑 개수가 기준 개수 이상인 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역 정렬(global alignment)을 수행하게 된다(118). 이때 상기 전역 정렬의 결과 리드의 에러 개수가 기 설정된 최대 에러 허용치(MaxError)를 초과하는 경우에는 정렬 실패로, 그렇지 않은 경우에는 정렬에 성공한 것으로 판단된다(120).

[35] 이하에서는 상기 112 단계 내지 118 단계의 구체적인 과정을 상세히 설명한다.

[36]

[37] 리드로부터 복수 개의 단편 서열 생성(112)

[38]

[39] 본 단계는 본격적으로 리드의 정렬을 수행하기 위하여 리드로부터 복수 개의 작은 조각인 단편 서열을 생성하는 단계이다. 본 단계에서는, 상기 리드의 첫 번째 베이스부터 마지막 베이스까지 설정된 간격(shift size) 만큼 이동하면서 설정된 크기(fragment size)만큼 리드의 값을 읽음으로써 상기 단편 서열들을 생성하게 된다.

[40] 도 3은 상기 112 단계에서의 단편 서열 생성 과정을 설명하기 위한 도면이다. 도시된 실시예에서는 리드의 길이가 75bp(base pair), 리드의 최대 에러 허용치가 3bp, 단편 서열의 크기(fragment size)가 15bp, 이동 간격(shift size)가 4bp인 경우의 실시예를 나타낸 것이다. 즉, 리드의 첫 번째 베이스부터 4bp씩 오른쪽으로 이동하면서 단편 서열을 생성하게 된다. 다만, 도시된 실시예의 경우 단지 예시적인 것으로서, 예컨대 상기 이동 간격, 단편 서열의 크기 등은 리드의 길이, 리드의 최대 에러 허용치 등의 값을 고려하여 적절하게 정해질 수 있다. 다시 말해 본 발명의 권리범위는 특정한 단편 서열의 크기 및 이동 간격에 한정된 것은 아님에 유의한다.

[41]

[42] 생성된 단편 서열의 필터링 및 확장(114)

[43]

[44] 상기와 같은 과정을 거쳐 단편 서열들이 생성되면, 다음으로 생성된 단편 서열 중 참조 서열과 매칭되지 않는 단편 서열들을 제외하는 필터링 과정을 거쳐 후보 단편 서열 집합(sub-candidate)을 구성한다. 즉, 생성된 단편 서열들과 상기 참조 서열과의 일치 정합(exact matching)을 시도하고, 그 결과 불일치하는 베이스의 수가 기 설정된 허용치 이하인 단편 서열들(후보 단편 서열들)로 상기 후보 단편 서열 집합을 구성하게 된다. 이때, 상기 허용치가 0인 경우, 상기 후보 단편 서열 집합에는 상기 참조 서열과 일치 정합되는 단편 서열들만이 포함된다.

[45] 예를 들어, 도 3에 도시된 실시예에서, 상기 리드의 15번째, 34번째 및 61번째 자리에서 에러가 발생하였다고 가정하자(도면에서 점선으로 표시). 이 경우 상기 에러를 포함하는 단편 서열들(도면에서 회색으로 표시)의 경우에는 참조 서열과의 일치 정합이 되지 않으며, 에러의 영향을 받지 않는 17-31, 37-51, 41-55 및 45-59의 4개의 단편 서열들만이 참조 서열과 일치 정합된다. 따라서 이 경우 상기 후보 단편 서열 집합에는 상술한 4개의 단편 서열들만이 포함된다.

[46] 한편, 일반적으로 참조 서열(예를 들어 인간의 유전체)은 다수의 중복 시퀀스(repeat sequence)를 포함한다. 이러한 중복 시퀀스는 참조 서열의 여러 위치에 분포하며, 동일한 염기 서열을 반복적으로 포함하고 있기 때문에 일부 단편 서열들의 경우 참조 서열과의 맵핑 시 너무 많은 위치에서 일치 정합이 발생하게 된다. 그러나 이와 같은 중복 시퀀스로 인해 일부 단편 서열들에서 너무 많은 수의 맵핑이 생성될 경우 전체 정렬 알고리즘의 복잡도 및 정확도에 악영향을 끼치므로, 이 경우 적절한 방법을 이용하여 맵핑되는 위치의 중복수를 줄일 필요가 있다.

[47] 이를 위하여, 본 단계에서는 후보 단편 서열들 중 상기 참조 서열에서의 맵핑 중복수가 기 설정된 값(예를 들어 50개)을 초과할 경우 상기 맵핑 중복수가 상기 설정된 값 이하가 될 때까지 해당 단편 서열의 크기를 확장하는 단계를 더 포함할 수 있다.

[48] 구체적으로, 본 단계에서는 생성된 상기 후보 단편 서열들 각각의 상기 참조 서열에서의 맵핑 위치의 개수를 계산하고, 계산된 맵핑 중복수(해당 단편 서열의 참조 서열에서의 맵핑 위치의 개수)가 설정된 값을 초과하는 단편 서열을 선택한 뒤, 상기 참조 서열에서의 맵핑 중복수가 상기 설정된 값 이하가 될 때까지 선택된 단편 서열의 크기를 확장하게 된다. 이때 상기 선택된 단편 서열의 크기 확장은, 상기 선택된 단편 서열의 처음 또는 끝 부분에, 해당 위치에 대응되는 상기 리드에서의 베이스를 추가함으로써 수행될 수 있다.

[49] 이를 예를 들어 설명하면 다음과 같다. 예를 들어, 다음과 같이 리드로부터 단편 서열이 생성되었다고 가정하자.

[50]

[51] 리드: A TTGC CTCAGT

[52] 단편 서열: TTGC (리드에서 밑줄 친 부분)

[53]

[54] 만약 상기 단편 서열에 대한 맵핑 결과 참조 서열에서의 맵핑 중복수가 기준치인 50개를 초과하는 65개일 경우에는, 아래와 같이 상기 맵핑 중복수가 기준치 이하로 떨어질 때까지 상기 단편 서열의 길이를 1bp씩 확장하게 된다.

[55]

[56] TTGC (65 맵핑 위치)

[57] TTGCC (54 맵핑 위치)

[58] TTGCCT (27 맵핑 위치)

[59]

[60] 위의 예시의 경우, 리드를 참조하여 2개의 베이스를 추가하였을 경우 맵핑 중복수가 설정값 이하로 떨어지므로, 최종 단편 서열은 최초 생성된 값보다 2bp 확장된 TTGCCT가 된다. 한편, 전술한 다른 예에서와 마찬가지로, 상기 설정값 또한 참조 서열, 리드 및 단편 서열의 특성 등에 따라서 적절하게 정해질 수 있는 값으로서 본 발명은 특정한 중복수 설정값에 그 권리범위가 한정되는 것은 아님에 유의하여야 한다.

[61]

인간의 염기 서열을 대상으로 한 실험에서, 100만 개의 리드에서 단편 서열의 길이 15bp, 시프트 간격 4bp로 단편 서열들을 생성한 후 생성된 단편 서열들을 참조 서열에 매칭할 경우, 기준치를 50으로 할 때 총 15,547,856개의 단편 서열 중 약 77%의 단편 서열이 50개 이하의 맵핑을 가지는 것으로 나타났다. 즉, 실험 결과 기준치를 50으로 할 경우 77%의 단편 서열은 이를 그대로 활용할 수 있으며, 나머지 23%의 단편 서열들은 전술한 방법에 따른 단편 서열 확장이 필요한 것으로 나타났다.

[62]

[63] 참조 서열의 각 구간별 맵핑 개수 계산(116)

[64]

[65] 상술한 과정을 통하여 후보 단편 서열 집합(sub-candidate)이 구성되면, 원칙적으로 이들 후보 단편 서열 집합들의 상기 참조 서열 내에서의 맵핑 위치들을 이용하여 리드를 참조 서열에 맵핑하는 것이 가능하다. 그러나 이 경우 후보 단편 서열들의 각 맵핑 위치의 모든 조합들을 고려하여야 하기 때문에 리드의 맵핑을 위한 계산의 복잡도가 매우 높아지게 된다. 예를 들어, 후보 단편 서열 집합에 포함된 후보 단편 서열이 4개이고, 각각의 후보 단편 서열들의 참조 서열에서의 맵핑 위치의 개수가 각각 3, 6, 24, 49개일 경우, $21,168(=3 \times 6 \times 24 \times 49)$ 개의 조합을 모두 조사하여야 한다. 본 발명에서는 이와 같은 계산의 복잡도를 줄이기 위해 참조 서열을 복수 개의 구간으로 분할하고, 이 중 맵핑 가능성이 높은 구간에 대해서만 전역 정렬을 수행하는 방법을 이용하게 된다.

[66]

즉, 본 발명에서는 먼저 참조 서열을 동일한 크기를 가지는 복수 개의 구간으로 분할하고, 분할된 각 구간 별로 다음의 값을 계산한다.

[67]

[68] A: 해당 구간에 맵핑되는 후보 단편 서열의 총 개수(맵핑 개수)

[69] B: 해당 구간에 맵핑되는 상기 후보 단편 서열의 총 맵핑 길이

[70]

[71] 예를 들어, 도 3에 도시된 실시예에서 17-31의 단편 서열이 분할된 첫 번째 구간에 맵핑될 경우 해당 구간의 (A, B) 값은 (1, 15)가 된다(이때, 1은 해당 구간에 맵핑되는 후보 단편 서열의 총 개수, 15는 맵핑되는 후보 단편 서열의 총 맵핑 길이). 동일한 방식으로, 37-51의 단편 서열이 두 번째 구간에 맵핑될 경우 해당 구간의 (A, B) 값은 (1, 15)가 된다. 이후 41-55의 단편 서열이 다시 상기 두 번째 구간에 맵핑될 경우 해당 구간의 (A, B) 값은 (2, 19)로 업데이트되며 그 이유는 다음과 같다.

[72]

[73] 첫 번째 값 2: 해당 구간에 맵핑되는 후보 단편 서열의 총 개수

[74] 두 번째 값 19: 최초 맵핑된 37-51 및 이후에 맵핑된 41-55의 오버랩되는 구간을 고려한 총 맵핑 길이

[75]

[76] 맵핑 대상 구간 선택 및 전역 정렬(Global Alignment)(118)

[77]

[78] 상기와 같은 과정을 거쳐 각 구간별 맵핑 개수 및 맵핑 길이가 계산되면, 이 중 맵핑 개수가 설정된 기준 개수 이상인 구간을 맵핑 대상 구간으로 선택한다. 또한, 상기 기준 개수 이상이 구간이 복수 개인 경우에는, 상기 총 맵핑 개수가 기준 개수 이상인 구간 중 상기 총 맵핑 길이가 설정된 기준 길이 이상인 구간을 맵핑 대상 구간으로 선택할 수 있다. 이때 상기 기준 개수는 적어도 2 이상이어야 하며, 그 이유는 맵핑의 기본 단위가 단편 서열임을 고려할 때 단편 서열이 하나만 맵핑되는 구간은 리드가 맵핑될 가능성이 매우 낮기 때문이다. 상기 기준 길이의 상세한 사항에 대해서는 후술한다.

[79] 도 4는 본 발명의 일 실시예에 따른 맵핑 대상 구간 선택 과정을 예시한 도면이다. 도시된 바와 같이, 참조 서열이 구간 1부터 4까지 까지 4개의 구간으로 분할되고, 각 구간의 맵핑 개수 및 맵핑 길이가 아래와 같이 계산되었다고 가정하자.

[80]

[81] 구간 1 = (1,15)

[82] 구간 2 = (0, 0)

[83] 구간 3 = (2,23)

[84] 구간 4 = (2,27)

[85]

[86] 이때, 상기 기준 개수가 2개로, 기준 길이가 22로 설정될 경우 상기 기준 개수 및 기준 길이를 만족하는 구간은 구간 3 및 4가 되며, 이에 따라 본 단계에서는 상기

구간 3 및 4에 해당하는 구간을 맵핑 대상 구간으로 선택하게 된다. 이때, 만약 상기 기준 개수 및 기준 길이를 만족하는 구간이 복수 개일 경우에는 해당하는 모든 구간이 맵핑 대상 구간이 되며, 맵핑 대상 구간에 포함된 복수 개의 구간 모두에서 전역 정렬이 수행된다. 이 경우에는 정렬 속도를 높이기 위하여 맵핑 대상 구간에 포함된 각 구간들의 맵핑 개수 또는 맵핑 길이를 서로 비교하고, 맵핑 개수 또는 맵핑 길이가 큰 구간부터 순차적으로 전역 정렬을 수행할 수 있다. 맵핑 개수 또는 맵핑 길이가 클 경우 해당 구간에서 리드가 맵핑될 가능성이 높기 때문이다. 예를 들어, 상기 실시예에서 구간 3 및 구간 4의 맵핑 개수는 2로 동일하나, 맵핑 길이 값은 구간 4가 3에 비해 크므로, 이 경우 구간 4부터 전역 정렬을 수행할 수 있다.

[87] 이와 같이 맵핑 대상 구간이 선택되면, 다음으로 후보 단편 서열(sub-candidate) 중 해당 맵핑 대상 구간에 맵핑되는 후보 단편 서열들을 최종 후보 단편 서열(candidate)로 선정하고, 선정된 최종 후보 단편 서열들 각각의 맵핑 위치에서 리드에 대한 전역 정렬을 수행함으로써 리드에 대한 정렬을 완료한다.

[88] 예를 들어, 상기 도 4의 실시예에서 구간 4에 맵핑되는 후보 단편 서열들이 37-51, 41-55, 45-59의 3개라고 가정하면, 상기 3개의 후보 단편 서열들이 최종 후보가 되며, 이들의 해당 구간 내 맵핑 위치에서 리드의 전역 정렬을 수행하게 된다.

[89] 한편, 상기 최종 후보 단편 서열에 대한 전역 정렬 시에는, 전역 정렬에 소요되는 시간을 줄이기 위하여 한 번 전역 정렬을 수행한 참조 서열에서의 위치를 기억하여 두었다가 그와 가까운 위치에서는 반복하여 여러 번 전역 정렬이 수행되지 않도록 한다. 구체적으로, 본 단계에서는 상기 맵핑 대상 구간을 복수 개의 소구간으로 분할한 뒤, 전역 정렬이 수행된 소구간의 경우 이를 기록하여 두도록 구성된다. 이후, 해당 소구간에 대한 전역 정렬 시에는 상기 기록된 정보를 이용하여 해당 소구간에서 전역 정렬이 기 수행되었는지의 여부를 판단하고, 상기 판단 결과 전역 정렬이 기 수행되지 않은 경우에만 전역 정렬을 수행하게 된다.

[90] 이를 예를 들어 설명하면 도 5와 같다. 도시된 바와 같이 맵핑 대상 구간이 5개의 소구간으로 나뉘어지고, 상기 3개의 최종 후보 중 37-51, 41-55는 두번째 소구간에, 45-59는 4번째 소구간에 맵핑된다고 가정하자. 이 경우 만약 37-51의 단편 서열에 대하여 2번째 소구간에서 전역 정렬이 수행될 경우, 그 결과에 관계 없이 동일한 소구간에 속한 41-55에 대해서는 전역 정렬이 수행되지 않으며, 이는 그 반대의 경우도 마찬가지이다. 따라서 도시된 실시예의 경우, 전역 정렬은 37-51/45-59 또는 41-55/45-59의 조합에 대해서만 실시되게 된다. 본 발명에서와 같이 참조 서열 전체가 아닌 맵핑 대상 구간 내에서만 전역 정렬이 수행된다고 하더라도 전역 정렬을 위해서는 시간이 많이 소요되므로, 이와 같은 과정을 통할 경우 전역 정렬에 소요되는 시간을 줄일 수 있다.

[91]

[92] 기준 길이 계산

[93]

[94] 상기 실시예에서, 기준 길이는 다음과 같은 방식으로 계산될 수 있다.

[95] 먼저, f 를 단편 서열의 크기, s 를 단편 서열을 생성하기 위한 리드 내에서의 이동 간격, L 을 리드의 길이, e 를 리드에서 허용되는 최대 에러의 개수, H 를 기준 길이라 할 때, 리드에서 에러의 영향을 받지 않는 영역의 길이 T 는 아래 수식과 같이 구할 수 있다.

[96]

[97] $T = L - f * e - s$

[98]

[99] 이때, L 과 e 는 본 발명의 수행 시 미리 결정되어 있는 값이므로, f , s 값에 따라 T 가 결정된다. 즉, f 와 s 값을 어떻게 변화시키느냐에 따라 알고리즘의 성능이 변화하게 된다.

[100] 먼저, H 값을 결정할 때는 아래의 두 가지 조건을 고려한다. 이 중 필수 조건은 반드시 충족해야 하며, 추가 조건은 가능한 경우에 고려한다.

[101] - 필수 조건: 맵핑의 기본 단위가 단편 서열이기 때문에, 기준 길이는 아무리 작더라도 적어도 오버랩되는 2개 이상의 단편 서열을 포함할 수 있는 크기여야 한다. 만약 도 2에서와 같이 $f=15$, $s=4$ 인 경우 오버랩되는 2개의 단편 서열들의 최소 길이는 $15+4=19$ 가 되므로, 적어도 H 값은 19 이상이어야 한다. 또한, 상기 H 값은 적어도 2개의 단편 서열이 포함되도록 설정되어야 하므로 최소한 $f+s$ 보다는 크거나 같아야 한다. 후술할 바와 같이, f 값은 최소 15 이상이어야 하므로, s 값을 그 최소값인 1로 가정할 경우 H 는 최소한 $16 (=15 + 1)$ 이상의 값이 된다.

[102] - 추가 조건: 이상적인 상황을 가정했을 때, $H=T$ 로 설정하고 T 이상의 시퀀스가 맵핑된 구간을 찾으면 주어진 에러에 대한 모든 맵핑을 찾을 수 있다. 그러나 전술한 바와 같이 참조 서열 자체에 중복이 많을 경우 상황에 따라 단편 서열의 길이를 확장하여야 할 경우가 발생할 수 있다. 따라서 이를 고려하여 H 값을 정할 때는 T 보다 약간 작은 Ts 를 사용하는 것이 맵핑률 측면에서 유리하다. 만약 $H=T$ 로 가정할 경우, $H=L - f * e - s$ 가 되며, 이 중 e 를 최소값인 1로 가정할 경우(e 가 0인 경우는 참조 서열과 일치 정합되는 경우이므로 전술한 104 단계에서 맵핑이 완료됨), $H=L - f - s$ 가 된다. 이 값이 기준 길이의 최대값이 된다. 만약 $L = 75\text{bp}$, $f = 15\text{bp}$, $s = 1$ 로 가정할 경우, H 의 최대값은 $75 - 15 - 1 = 59$ 가 된다.

[103] 정리하면, 상기 H 값은 다음의 범위를 만족하여야 한다.

[104]

[105] $f + s \leq H \leq L - (f + s)$

[106]

[107] 다음으로, f 값은 아래 두 가지 조건을 만족하는 값 중에서 큰 값을 고른다. 역시

필수 조건은 반드시 충족해야 하며, 추가 조건은 가능한 경우에 고려한다.

[108] - 필수 조건: f 는 15 이상이어야 하며, 그 이유는 단편 서열의 길이가 14 이하일 경우 참조 서열 내에서의 맵핑 위치의 개수가 급격히 증가하기 때문이다.

[109] 아래의 표 1은 단편 서열 길이에 따른 인간 유전체 내에서의 단편 서열의 평균 등장 빈도를 나타낸 것이다.

[110]

[111] 표 1

[Table 1]

단편 서열의 길이	평균 등장 빈도
10	2,726.1919
11	681.9731
12	170.9185
13	42.7099
14	10.6470
15	2.6617
16	0.6654
17	0.1664

[112]

[113]상기 표에서 알 수 있는 바와 같이, 단편 서열의 길이가 14 이하일 경우에는 단편 서열 별 빈도가 10 이상이나, 15일 경우에는 3 이하로 감소하는 것을 알 수 있다. 즉, 단편 서열의 길이를 15 이상으로 구성할 경우 14 이하로 구성할 경우에 비해 단편 서열의 중복을 대폭 감소시킬 수 있다.

[114] - 추가 조건: $f = L/(e+2)$ 을 만족하여야 하며, 이는 T 의 길이를 단편 서열 2개의 크기 이상으로 보장하기 위함이다.

[115] 예를 들어, $L=100$, $e=4$ 일 때 f 는 16 이하의 값을 가져야 한다.

[116]

[117] 위의 조건을 정리하여, f 와 s , H 를 결정하는 방법을 정리하면 다음과 같다.

[118] - s 는 4로 고정한 뒤, f 와 H 를 결정한다.

[119] - $15 \leq f \leq L/(e+2)$ 범위 내에서 가장 큰 값을 f 로 결정한다. (단, 반드시 $f = 15$)

[120] - H 는 아래 식을 이용해 결정한다.

[121] $H = L - f * e - 2s$ 또는 $H = f + s$ 에서 계산되는 값 중 큰 값

[122] (이때, H 는 기준 길이, L 은 리드의 길이, f 는 단편 서열의 길이, e 는 리드의 최대 에러 개수, s 는 각 단편 서열들의 이동 간격)

[123]

[124] 예 1) $L = 75$, $e = 3$ 일 때,

- [125] $f = 15 \sim 15$ 이므로 15,
 [126] $s = 4$,
 [127] $H = 75 - 3 \times 15 - 2 \times 4 = 22$ 가 된다.
 [128]
 [129] 예 2) $L = 100$, $e = 4$ 일 때,
 [130] $f = 15 \sim 16$ 이므로 16,
 [131] $s = 4$,
 [132] $H = 100 - 4 \times 16 - 2 \times 4 = 36 - 8 = 28$ 가 된다.
 [133]
 [134] 예 3) $L = 75$, $e = 4$ 일 때
 [135] $f = 15 \sim 12$ 이지만, 15 이상이어야 하므로 15,
 [136] $s = 4$,
 [137] $H = 75 - 4 \times 15 - 2 \times 4 = 15 - 8 = 7$ 이지만, $f + s = 19$ 이므로, 결과적으로 $H = 19$ 가 된다.
 [138]
 [139] 도 6은 본 발명의 일 실시예에 따른 염기 서열 정렬 시스템(600)의 블록도이다. 본 발명의 일 실시예에 따른 염기 서열 정렬 시스템(600)은 전술한 염기 서열 정렬 방법을 수행하기 위한 장치로서, 단편 서열 생성부(602), 필터링부(604), 맵핑 개수 계산부(606), 정렬부(608) 및 단편 서열 확장부(610)를 포함한다.
 [140] 단편 서열 생성부(602)는 게놈 시퀀서로부터 얻어진 리드로부터 복수 개의 단편(fragment) 서열들을 생성한다. 전술한 바와 같이, 단편 서열 생성부(602)는 상기 리드의 첫 번째 베이스부터 설정된 간격만큼 이동하면서 설정된 크기만큼 상기 리드의 값을 읽음으로써 상기 단편 서열들을 생성하게 된다.
 [141] 필터링부(604)는 생성된 상기 복수 개의 단편 서열들 중 상기 참조 서열과 매칭되는 단편 서열만을 포함하는 후보 단편 서열 집합을 구성한다. 이때, 상기 참조 서열과 매칭되는 단편 서열은, 상기 참조 서열과의 일치 정합(exact matching) 결과 불일치하는 베이스의 수가 설정된 개수 이하인 단편 서열을 의미한다.
 [142] 맵핑 개수 계산부(606)는 상기 참조 서열을 복수 개의 구간으로 분할하고, 상기 각 구간 별로 상기 후보 단편 서열들의 맵핑 위치 및 각 구간 별 상기 후보 단편 서열들의 총 맵핑 개수를 계산한다.
 [143] 정렬부(608)는 맵핑 개수 계산부(606)에 의하여 분할된 구간 중 계산된 상기 총 맵핑 개수가 기준 개수 이상인 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역 정렬을 수행한다. 구체적으로, 정렬부(608)는 상기 후보 단편 서열 중 상기 선택된 구간에 맵핑되는 후보 단편 서열의 상기 참조 서열 내에서의 맵핑 위치에 기초하여 상기 리드에 대한 전역 정렬을 수행한다.
 [144] 또한 정렬부(608)는 상기 선택된 구간(맵핑 대상 구간)을 복수 개의 소구간으로 분할하고, 상기 전역 정렬을 수행하려는 상기 참조 서열 내 위치가 속한

소구간에서 전역 정렬이 기 수행되었는지의 여부를 판단하며, 상기 판단 결과 전역 정렬이 기 수행되지 않은 경우에만 상기 전역 정렬을 수행함으로써 불필요한 전역 정렬의 횟수를 감소시키도록 구성될 수 있다.

[145] 단편 서열 확장부(610)는 상기 필터링부(604)에서 생성된 상기 후보 단편 서열들 각각의 상기 참조 서열에서의 맵핑 중복수를 계산하고, 계산된 상기 맵핑 중복수가 설정된 값을 초과하는 단편 서열을 선택하며, 상기 참조 서열에서의 맵핑 중복수가 상기 설정된 값 이하가 될 때까지 선택된 단편 서열의 크기를 확장한다. 이때, 단편 서열 확장부(610)는 상기 선택된 단편 서열의 처음 또는 끝 부분에, 해당 위치에 대응되는 상기 리드에서의 베이스를 추가함으로써 상기 확장을 수행한다.

[146] 한편, 본 발명의 실시예는 본 명세서에서 기술한 방법들을 컴퓨터상에서 수행하기 위한 프로그램을 포함하는 컴퓨터 판독 가능 기록매체를 포함할 수 있다. 상기 컴퓨터 판독 가능 기록매체는 프로그램 명령, 로컬 데이터 파일, 로컬 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야에서 통상의 지식을 가진 자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD와 같은 광 기록 매체, 플로피 디스크와 같은 자기-광 매체, 및 롬, 램, 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

[147] 이상에서 대표적인 실시예를 통하여 본 발명에 대하여 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다.

[148] 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

[149]

[150] <부호의 설명>

[151] 600: 염기 서열 정렬 시스템

[152] 602: 단편 서열 생성부

[153] 604: 필터링부

[154] 606: 맵핑 개수 계산부

[155] 608: 정렬부

[156] 610: 단편 서열 확장부

청구범위

- [청구항 1] 리드로부터 복수 개의 단편(fragment) 서열들을 생성하는 단편 서열 생성부;
 생성된 상기 복수 개의 단편 서열들 중 참조 서열과 매칭되는 단편 서열만을 포함하는 후보 단편 서열 집합을 구성하는 필터링부;
 상기 참조 서열을 복수 개의 구간으로 분할하고, 상기 각 구간 별 상기 후보 단편 서열들의 총 맵핑 개수를 계산하는 맵핑 개수 계산부; 및
 계산된 상기 총 맵핑 개수가 기준 개수 이상인 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역 정렬(global alignment)을 수행하는 정렬부를 포함하는 염기 서열 정렬 시스템.
- [청구항 2] 청구항 1에 있어서,
 상기 단편 서열 생성부는, 상기 리드의 첫 번째 베이스부터 설정된 간격만큼 이동하면서 설정된 크기만큼 상기 리드의 값을 읽음으로써 상기 단편 서열들을 생성하는, 염기 서열 정렬 시스템.
- [청구항 3] 청구항 1에 있어서,
 상기 참조 서열과 매칭되는 단편 서열은, 상기 참조 서열과의 일치 정합(exact matching) 결과 불일치하는 베이스의 수가 설정된 개수 이하인 단편 서열인, 염기 서열 정렬 시스템.
- [청구항 4] 청구항 1에 있어서,
 상기 후보 단편 서열들 각각의 상기 참조 서열에서의 맵핑 중복수를 계산하고, 계산된 상기 맵핑 중복수가 설정된 값을 초과하는 단편 서열을 선택하며, 상기 참조 서열에서의 맵핑 위치의 개수가 상기 설정된 값 이하가 될 때까지 선택된 단편 서열의 크기를 확장하는 단편 서열 확장부를 더 포함하는, 염기 서열 정렬 시스템.
- [청구항 5] 청구항 4에 있어서,
 상기 단편 서열 확장부는, 상기 선택된 단편 서열의 처음 또는 끝 부분에, 해당 위치에 대응되는 상기 리드에서의 베이스를 추가하는, 염기 서열 정렬 시스템.
- [청구항 6] 청구항 1에 있어서,
 상기 정렬부는, 상기 후보 단편 서열 중 상기 선택된 구간에 맵핑되는 후보 단편 서열을 선택하고, 선택된 각각의 후보 단편 서열의 상기 참조 서열 내에서의 맵핑 위치에서 상기 리드에 대한 전역 정렬을 수행하는, 염기 서열 정렬 시스템.
- [청구항 7] 청구항 6에 있어서,
 상기 정렬부는, 상기 선택된 구간을 복수 개의 소구간으로

분할하고, 상기 전역 정렬을 수행하려는 상기 참조 서열 내 위치가 속한 소구간에서 전역 정렬이 기 수행되었는지의 여부를 판단하며, 상기 판단 결과 전역 정렬이 기 수행되지 않은 경우에만 상기 전역 정렬을 수행하는, 염기 서열 정렬 시스템.

[청구항 8]

청구항 1에 있어서,
상기 맵핑 개수 계산부는, 상기 총 맵핑 개수와 함께 상기 각 구간 별 상기 후보 단편 서열들의 총 맵핑 길이를 계산하며,
상기 정렬부는, 상기 총 맵핑 개수가 기준 개수 이상인 구간 중 상기 총 맵핑 길이가 설정된 기준 길이 이상이 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역 정렬을 수행하는, 염기 서열 정렬 시스템.

[청구항 9]

청구항 8에 있어서,
상기 정렬부는, 상기 선택된 구간이 복수 개인 경우, 각 구간 별 총 맵핑 개수 또는 총 맵핑 길이에 따라 순차적으로 상기 리드에 대한 전역 정렬을 수행하는, 염기 서열 정렬 시스템.

[청구항 10]

청구항 8에 있어서,
상기 기준 개수는 적어도 2 이상인, 염기 서열 정렬 시스템.

[청구항 11]

청구항 8에 있어서,
상기 기준 길이는 다음의 수학적식
$$H = L - f * e - 2s$$

(이때, H는 기준 길이, L은 리드의 길이, f는 단편 서열의 길이, e는 리드의 최대 에러 개수, s는 각 단편 서열들의 이동 간격)

또는, $H = f + s$

에서 계산되는 값 중 큰 값인, 염기 서열 정렬 시스템.

[청구항 12]

청구항 11에 있어서,
상기 기준 길이는 다음의 수학적식

$$f + s \leq H \leq L - (f + s)$$

을 만족하는, 염기 서열 정렬 시스템.

[청구항 13]

청구항 8에 있어서,
상기 기준 길이는 16 이상 59 이하인, 염기 서열 정렬 시스템.

[청구항 14]

단편 서열 생성부에서, 리드로부터 복수 개의 단편(fragment) 서열들을 생성하는 단계;
필터링부에서, 생성된 상기 복수 개의 단편 서열들 중 참조 서열과 매칭되는 단편 서열만을 포함하는 후보 단편 서열 집합을 구성하는 단계;
맵핑 개수 계산부에서, 상기 참조 서열을 복수 개의 구간으로 분할하고, 상기 각 구간 별로 상기 후보 단편 서열들의 총 맵핑 개수를 계산하는 단계; 및

정렬부에서, 계산된 상기 총 맵핑 개수가 기준 개수 이상인 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역 정렬(global alignment)을 수행하는 단계를 포함하며, 상기 참조 서열과 매칭되는 단편 서열은, 상기 참조 서열과의 일치 정합(exact matching) 결과 불일치하는 베이스의 수가 설정된 개수 이하인 단편 서열인, 염기 서열 정렬 방법.

[청구항 15]

청구항 14에 있어서, 상기 단편 서열들을 생성하는 단계는, 상기 리드의 첫 번째 베이스부터 설정된 간격만큼 이동하면서 설정된 크기만큼 상기 리드의 값을 읽음으로써 상기 단편 서열들을 생성하는, 염기 서열 정렬 방법.

[청구항 16]

청구항 14에 있어서, 상기 후보 단편 서열 집합을 구성하는 단계는, 단편 서열 확장부에서, 생성된 상기 후보 단편 서열들 각각의 상기 참조 서열에서의 맵핑 중복수를 계산하는 단계; 상기 단편 서열 확장부에서, 계산된 상기 맵핑 중복수가 설정된 값을 초과하는 단편 서열을 선택하는 단계; 및 상기 단편 서열 확장부에서, 상기 참조 서열에서의 맵핑 중복수가 상기 설정된 값 이하가 될 때까지 선택된 단편 서열의 크기를 확장하는 단계를 더 포함하며, 상기 선택된 단편 서열의 크기를 확장하는 단계는, 상기 선택된 단편 서열의 처음 또는 끝 부분에, 해당 위치에 대응되는 상기 리드에서의 베이스를 추가하도록 구성되는, 염기 서열 정렬 방법.

[청구항 17]

청구항 14에 있어서, 상기 전역 정렬을 수행하는 단계는, 상기 후보 단편 서열 중 상기 선택된 구간에 맵핑되는 후보 단편 서열을 선택하고, 선택된 각각의 후보 단편 서열의 상기 참조 서열 내에서의 맵핑 위치에서 상기 리드에 대한 전역 정렬을 수행하되, 상기 전역 정렬을 수행하는 단계는, 상기 선택된 구간을 복수 개의 소구간으로 분할하는 단계; 및 상기 전역 정렬을 수행하려는 상기 참조 서열 내 위치가 속한 소구간에서 전역 정렬이 기 수행되었는지의 여부를 판단하는 단계를 더 포함하며, 상기 판단 결과 전역 정렬이 기 수행되지 않은 경우에만 상기 전역 정렬을 수행하는, 염기 서열 정렬 방법.

[청구항 18]

청구항 14에 있어서, 상기 총 맵핑 개수를 계산하는 단계는, 상기 각 구간 별 상기 후보 단편 서열들의 총 맵핑 길이를 계산하는 단계를 더 포함하며, 상기 전역 정렬을 수행하는 단계는, 상기 총 맵핑 개수가 기준 개수

이상인 구간 중 상기 총 맵핑 길이가 설정된 기준 길이 이상이 구간을 선택하고, 선택된 구간에 대하여 상기 리드에 대한 전역 정렬을 수행하는, 염기 서열 정렬 방법.

[청구항 19]

청구항 18에 있어서,

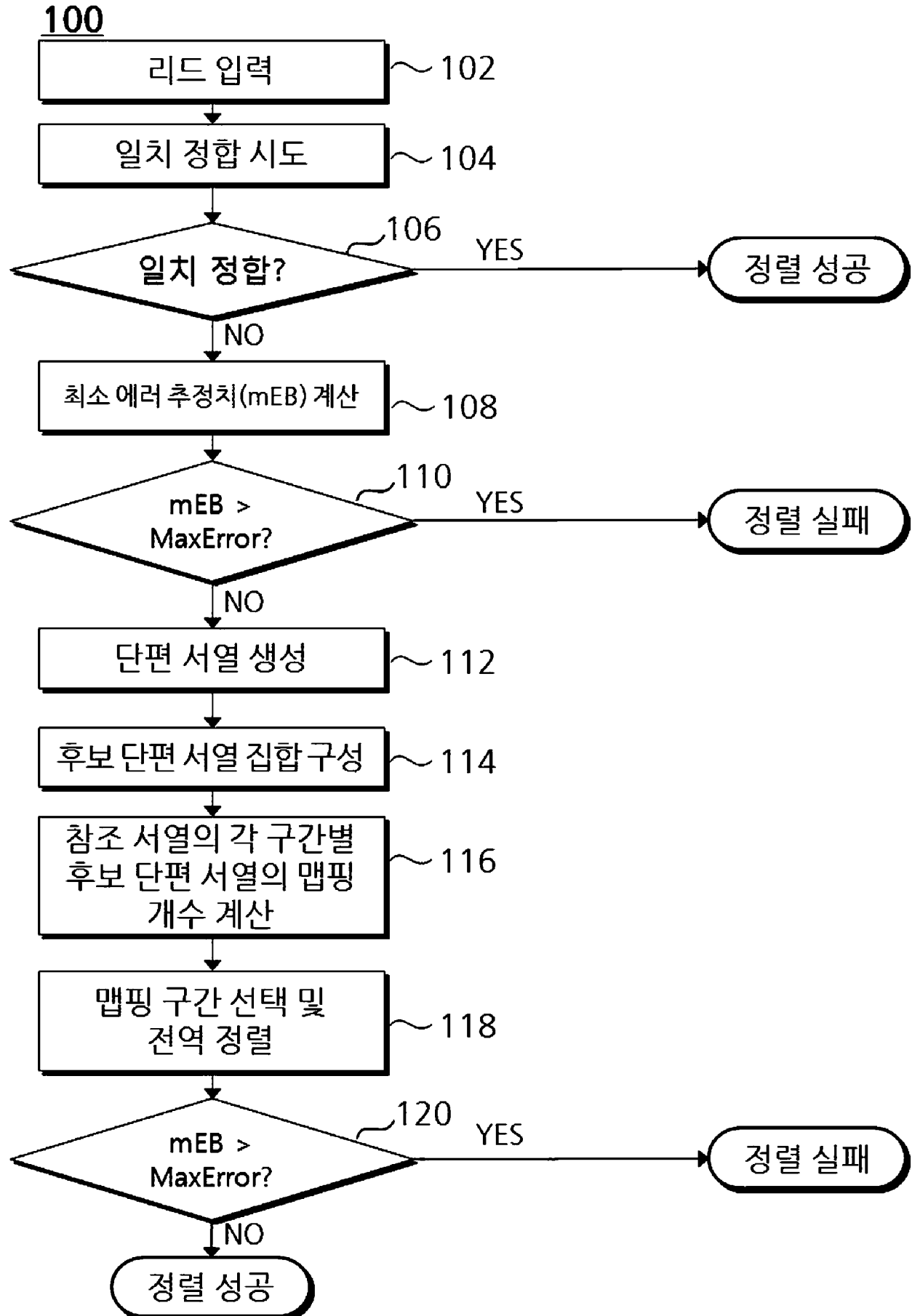
상기 전역 정렬을 수행하는 단계는, 상기 선택된 구간이 복수 개인 경우, 각 구간 별 총 맵핑 개수 또는 총 맵핑 길이에 따라 순차적으로 상기 리드에 대한 전역 정렬을 수행하는, 염기 서열 정렬 방법.

[청구항 20]

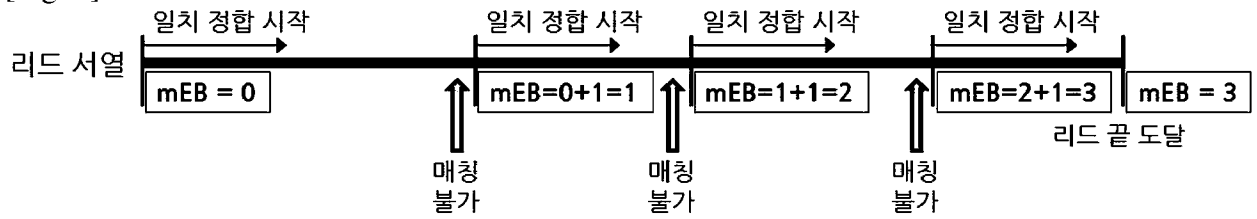
청구항 18에 있어서,

상기 기준 길이는 16 이상 59 이하인, 염기 서열 정렬 방법.

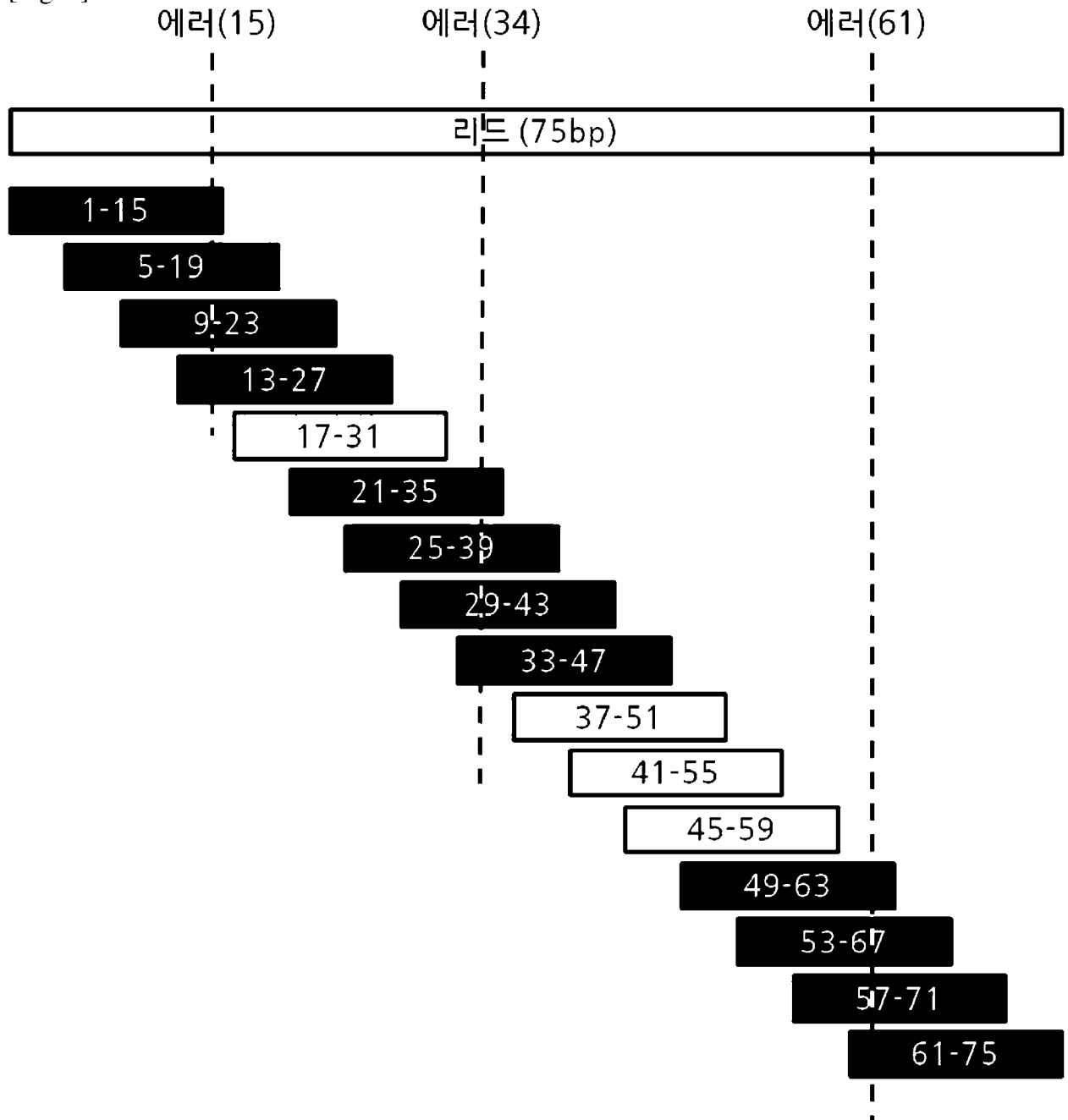
[Fig. 1]



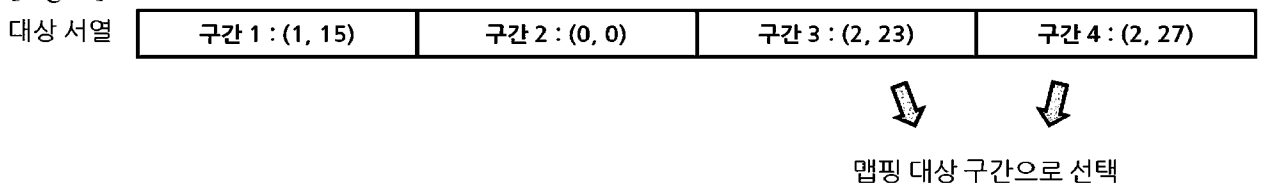
[Fig. 2]



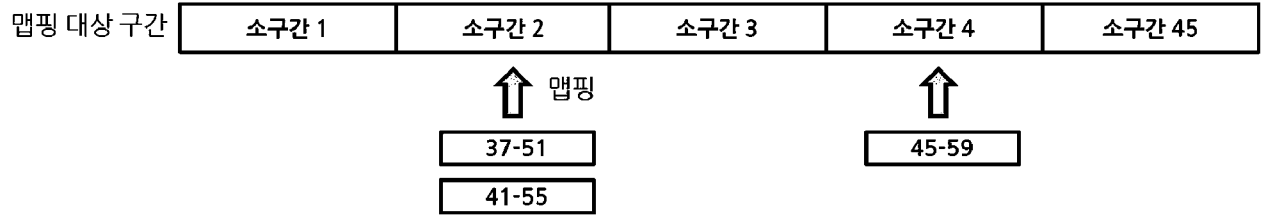
[Fig. 3]



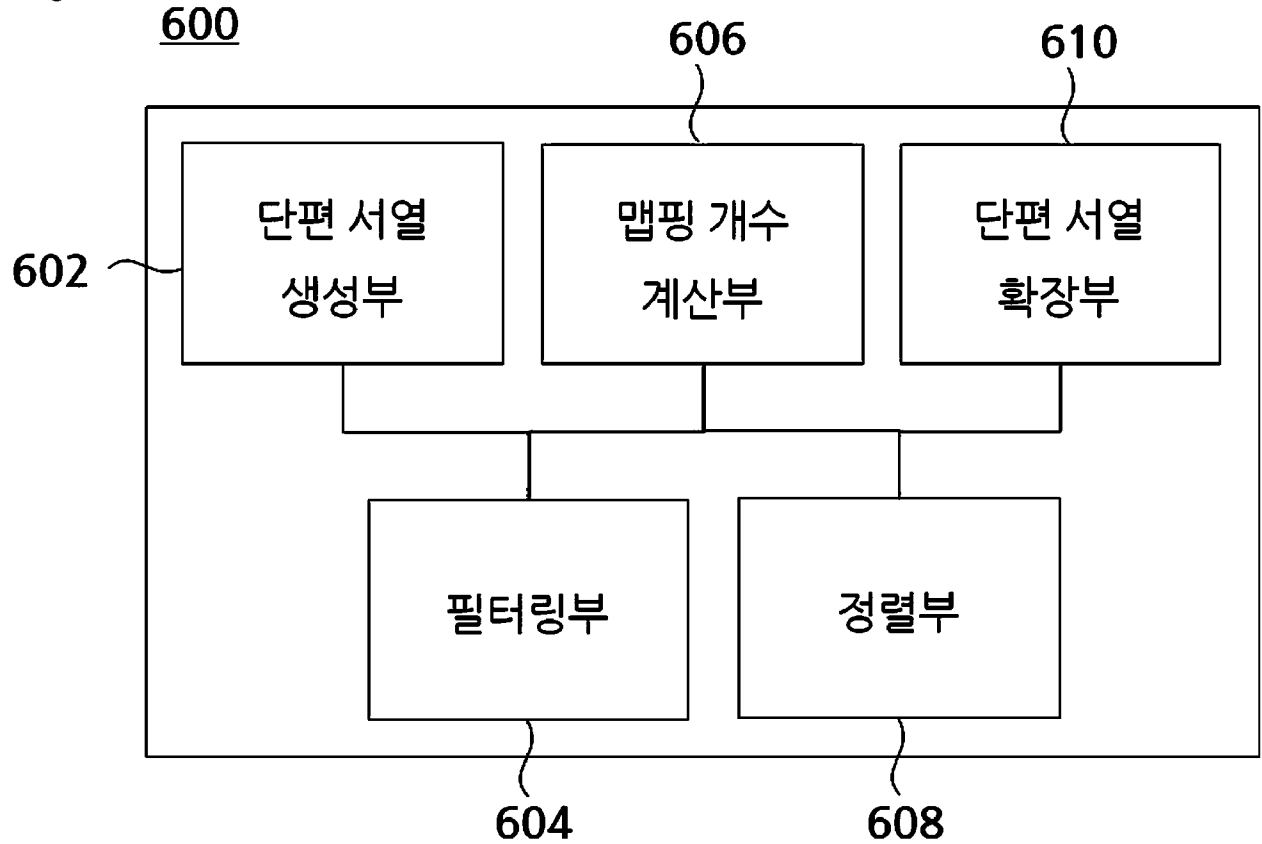
[Fig. 4]



[Fig. 5]



[Fig. 6]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2013/007276**A. CLASSIFICATION OF SUBJECT MATTER*****C12Q 1/68(2006.01)i, G06F 19/22(2011.01)i***

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12Q 1/68; G06F 19/22

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean Utility models and applications for Utility models: IPC as above

Japanese Utility models and applications for Utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & Keywords: nucleotide sequence recombination, read, seed, mapping, overlap, global alignment, filtering

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	LI et al., "Mapping short DNA sequencing reads and calling variants using mapping quality scores" Genome Research, vol. 18, no. 11, pp. 1851-1858 (2008) See abstract and pages 1853, 1855-1857.	1-20
A	RIZK et al., "GASSST: global alignment short sequence search tool" Bioinformatics, vol. 26, no. 20, pp. 2534-2540 (2010) See abstract and pages 2537-2539.	1-20
A	LANGMEAD et al., "Fast gapped-read alignment with Bowtie 2" Nature Methods, vol. 9, no. 4, pp. 357-359 (April 2012) See abstract, supplementary figure 1 and supplementary note.	1-20
A	PARK, Tae Won et al., "Method for Time Efficient Processing of Next Generation Sequencing Reads Using Hitting Set Problem" Journal of KISSE: Computer Systems and Theory, vol. 38, no. 6, pp. 275-280 (2011) See abstract and pages 278-279.	1-20
A	LI et al., "A survey of sequence alignment algorithms for next-generation sequencing" Briefings in Bioinformatics, vol. 11, no. 5, pp. 473-483 (2010) See the entire document.	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

29 OCTOBER 2013 (29.10.2013)

Date of mailing of the international search report

01 NOVEMBER 2013 (01.11.2013)

Name and mailing address of the ISA/KR



Korean Intellectual Property Office
Government Complex-Daejeon, 189 Seonsa-ro, Daejeon 302-701,
Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2013/007276

Patent document
cited in search report

Publication
date

Patent family
member

Publication
date

NONE

A. 발명이 속하는 기술분류(국제특허분류(IPC)) C12Q 1/68(2006.01)i, G06F 19/22(2011.01)i		
B. 조사된 분야 조사된 최소문헌(국제특허분류를 기재) C12Q 1/68; G06F 19/22 조사된 기술분야에 속하는 최소문헌 이외의 문헌 한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC 일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC 국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우)) eKOMPASS(특허청 내부 검색시스템) & 키워드: 염기서열 재조합, 리드, 시드, 맵핑, 중복, 전역정렬, 필터링		
C. 관련 문헌		
카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	LI 외, 'Mapping short DNA sequencing reads and calling variants using mapping quality scores' Genome Research, Vol.18, No.11, pp.1851-1858 (2008) 요약 및 페이지 1853, 1855-1857 참조.	1-20
A	RIZK 외, 'GASSST: global alignment short sequence search tool' Bioinformatics, Vol.26, No.20, pp.2534-2540 (2010) 요약 및 페이지 2537-2539 참조.	1-20
A	LANGMEAD 외, 'Fast gapped-read alignment with Bowtie 2' Nature Methods, Vol.9, No.4, pp.357-359 (2012.04) 요약, supplementary figure 1 및 supplementary note 참조.	1-20
A	박태원 외, 'Hitting Set 문제를 이용한 차세대 염기서열 분석 리드의 처리 속도 개선 방법' 정보과학회논문지: 시스템 및 이론, 제38권, 제6호, 페이지 275-280 (2011) 요약 및 페이지 278-279 참조.	1-20
A	LI 외, 'A survey of sequence alignment algorithms for next-generation sequencing' Briefings in Bioinformatics, Vol.11, No.5, pp.473-483 (2010) 전체 문헌 참조.	1-20
☐ 추가 문헌이 C(계속)에 기재되어 있습니다. ☒ 대응특허에 관한 별지를 참조하십시오.		
* 인용된 문헌의 특별 카테고리: "A" 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌 "E" 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌 "L" 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌 "O" 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌 "P" 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌 "T" 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌 "X" 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다. "Y" 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다. "&" 동일한 대응특허문헌에 속하는 문헌		
국제조사의 실제 완료일 2013년 10월 29일 (29.10.2013)		국제조사보고서 발송일 2013년 11월 01일 (01.11.2013)
ISA/KR의 명칭 및 우편주소  대한민국 특허청 (302-701) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사) 팩스 번호 +82-42-472-7140		심사관 김승범 전화번호 +82-42-481-3371 

국제조사보고서에서
인용된 특허문헌

공개일

대응특허문헌

공개일

없음