(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0025465
(43) 공개일자 2015년03월10일

(51) 국제특허분류(Int. Cl.)
G06F 19/00 (2011.01) *C12N 15/11* (2006.01)
(21) 출원번호 10-2013-0103383
(22) 출원일자 2013년08월29일
심사청구일자 2013년08월29일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
박준홍
서울 종로구 사직로8길 24, 1401호 (내수동, 경희궁의아침2단지)
양지훈
경기 광명시 오리로921번길 10, 201호 (광명동, 청원빌리지)
신승욱
서울 관악구 호암로20길 21, G 109호 (신림동)
(74) 대리인
이덕록

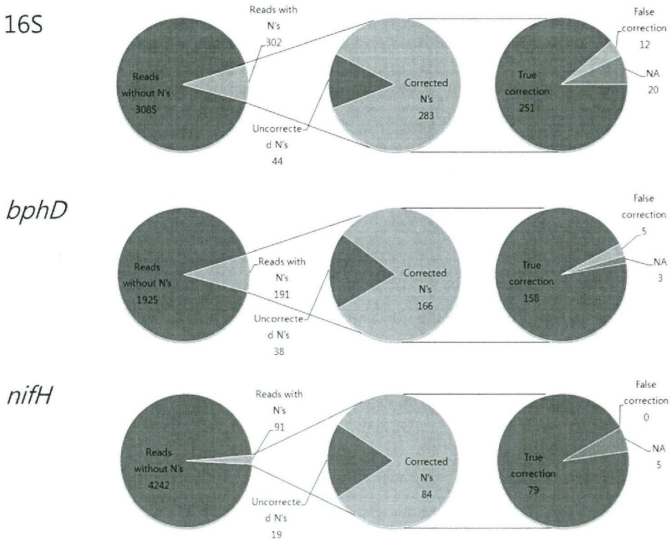
전체 청구항 수 : 총 3 항

(54) 발명의 명칭 **컴퓨터를 기반으로 하는 모호한 시퀀스 보정 방법**

(57) 요 약

본 발명은 컴퓨터를 기반으로 하여 모호한 시퀀스를 검측 및 보정하는 방법에 관한 것으로 인공 유전체 DNA를 증폭하고 파이로시퀀싱하는 단계와; 상기 단계에서 얻은 시퀀스를 표준염기서열과 비교하는 단계와; 상기 단계에서 얻은 DNA의 시퀀스 모티프에서 모호한 시퀀스의 발생 패턴을 확인하는 단계와; 본 발명 프로그램을 이용하여 모호한 시퀀스를 보정하고 확인하는 단계를 통하여 결과적으로 PCR 및 파이로시퀀싱 결과물의 오류를 저감시킬 수 있는 뛰어난 효과가 있다.

대 표 도 - 도2



이 발명을 지원한 국가연구개발사업
과제고유번호 2012000550030
부처명 환경부
연구관리전문기관 한국환경산업기술원
연구사업명 토양지하수오염방지기술개발사업
연구과제명 토양지하수오염 생물학적 복원공법 선정 및 최적화 지원을 위한 RNA-메타지노믹스 기반의 BT/IT/ET 융합기술 개발연구
기 여 율 1/1
주관기관 연세대학교 산학협력단
연구기간 2012.07.01 ~ 2013.06.30

특허청구의 범위

청구항 1

컴퓨터를 기반으로 하는 모호한 시퀀스 보정 방법에 있어서, 프로세서, 데이터 저장 시스템, 입력장치 및 출력장치를 포함하는 프로그램된 컴퓨터를 이용하여

(a)타겟 유전자 시퀀스를 포함하는 데이터를 입력장치를 통해서 프로그램된 컴퓨터에 입력하는 단계;

(b)프로세서를 이용하여 모호한 시퀀스를 호모폴리머 시퀀스 정보와 비교하여 보정하는 단계;

(c)상기 보정한 결과를 출력장치에 출력하는 단계

를 포함하는 것을 특징으로 하는 컴퓨터를 기반으로 하는 모호한 시퀀스 보정 방법.

청구항 2

제1항에 있어서, 상기 호모폴리머는 모호한 시퀀스의 상부에 동일한 염기가 반복되는 형태로 존재하는 것을 특징으로 하는 모호한 시퀀스 보정 방법.

청구항 3

미생물의 DNA를 추출하여 인공유전체를 제조하는 단계:

상기 인공유전체 DNA를 증폭하는 단계;

상기 DNA를 정제하여 파이로시퀀싱하는 단계;

상기 단계에서 얻은 시퀀스에서 모호한 시퀀스(N's) 상부에 동일한 염기가 반복되는 호모폴리머를 찾는 단계;

상기 모호한 시퀀스를 반복되는 동일한 염기로 결정하는 단계

로 이루어지는 것이 특징인 파이로시퀀싱의 모호한 시퀀스 보정 방법.

명 세 서

기 술 분 야

[0001] 본 발명은 컴퓨터를 기반으로 하여 모호한 시퀀스를 검측 및 보정하는 방법에 관한 것이다.

배 경 기 술

[0002] 대량 병렬 454 파이로시퀀싱(massively parallel 454 pyrosequencing) 기법은 최근 개발되어 대중화된 방법으로 특히 16S rRNA 유전자의 파이로시퀀싱을 통해 배양기반의 방법 없이도 미생물 군집에서의 분자적 변화 및 분류체계를 평가할 수 있게 되었다. 파이로시퀀싱(pyrosequencing)은 유전체 시퀀싱에 사용되는 가장 보편적인 방법으로 높은 처리용량의 시퀀스 정보를 빠르고 저렴하게 획득할 수 있는 장점이 있다. 파이로시퀀싱 시스템에 관하여는 호모폴리머(homopolymer)의 정확한 길이 결정에 초점이 맞추어져 있으며 시퀀싱의 오류는 미생물 군집의 종 풍부도 계산에 영향을 미칠 수 있다. 파이로시퀀싱 단계에서 발생할 수 있는 오류는 4가지 유형으로 구분된다: 삽입(insertion), 결실(deletion), 미스매치(mismatch), 및 모호한 염기(ambiguous base). 모호한 시퀀스(N's)는 454 파이로시퀀싱에서 유의적으로 중요한 오류원인으로 모호한 시퀀스 오류의 수는 전체 오류수의 21%를 차지한다(Huse *et al.*,2007).

[0003] 파이로시퀀싱의 정확도 측정용 인공유전체 및 측정방법에 관하여는 대한민국 등록특허 제10-1259144호가 개시되어 있다.

[0004] 한편 컴퓨터를 이용하여 보존된 불변 펩티드 모티프를 확인하는 방법에 관하여는 대한민국 등록특허 제10-0780874호가 개시되어 있고 컴퓨터 시스템을 이용하여 잠재적인 에리트로포이에틴 유사물질 동정방법에 관하여

는 등록특허 제10-0458755호가 개시되어 있다. 그러나 상기 문헌 어디에도 컴퓨터를 이용하여 파이로시퀀싱을 보정하는 방법에 관하여는 개시된 바 없다.

발명의 내용

해결하려는 과제

[0005] 따라서 본 발명의 목적은 모호한 시퀀스의 검측 및 보정이 가능한 펄(Perl) 기반의 프로그램을 제공하는 데 있다.

과제의 해결 수단

[0006] 본 발명의 목적은 인공 유전체 DNA를 증폭하고 파이로시퀀싱하는 단계와; 상기 단계에서 얻은 시퀀스를 표준염기서열과 비교하는 단계와; 상기 단계에서 얻은 DNA의 시퀀스 모티프에서 모호한 시퀀스의 발생 패턴을 확인하는 단계와; 본 발명 프로그램을 이용하여 모호한 시퀀스를 보정하고 확인하는 단계를 통하여 달성하였다.

발명의 효과

[0007] 본 발명은 PCR 및 파이로시퀀싱 결과물의 오류를 저감시킬 수 있는 뛰어난 효과가 있다.

도면의 간단한 설명

[0008] 도 1은 모호한 시퀀스의 분포를 나타낸 그래프이다.

도 2는 본 발명 프로그램을 이용하여 모호한 시퀀스 보정 결과를 나타낸 다이어그램이다.]

도 3, 4 및 5는 본 발명 컴퓨터 프로그램의 스크립트이다.

발명을 실시하기 위한 구체적인 내용

[0009] 본 발명은 컴퓨터를 기반으로 하는 모호한 시퀀스의 보정 방법을 제공한다.

[0010] 본 발명은 바람직하게는 컴퓨터를 기반으로 하는 모호한 시퀀스 보정 방법에 있어서, 프로세서, 데이터 저장 시스템, 입력장치 및 출력장치를 포함하는 프로그램된 컴퓨터를 이용하여

[0011] (a)타겟 유전자 시퀀스를 포함하는 데이터를 입력장치를 통해서 프로그램된 컴퓨터에 입력하는 단계;

[0012] (b)프로세서를 이용하여 모호한 시퀀스를 호모폴리머 시퀀스 정보와 비교하여 보정하는 단계;

[0013] (c)상기 보정한 결과를 출력장치에 출력하는 단계로 구성된다.

[0014] 본 발명은 하드웨어나 소프트웨어 또는 그들의 조합을 이용할 수 있다. 그러나 바람직하기로 본 발명은 프로세서, 데이터 저장시스템(휘발성 및 비-휘발성 메모리 및/또는 저장요소를 포함), 적어도 하나의 입력장치 및 적어도 하나의 출력장치를 구비하는 프로그램 가능한 컴퓨터 상에서 실행되는 컴퓨터 프로그램을 이용할 수 있다. 프로그램 코드가 상술한 기능을 수행하고 출력정보를 생성하도록 입력데이터에 부여된다. 출력정보는 공지의 방식으로 하나 이상의 출력장치에 공급된다.

[0015] 본 발명의 프로그램은 컴퓨터 시스템과 소통하기 위해 바람직하게는 하이레벨 절차 프로그래밍 언어(high level procedural programming language) 또는 객체지향 프로그래밍 언어(object oriented programming language)를 이용할 수 있다. 그러나 프로그램은 원한다면 어셈블리어 또는 기계어를 이용할 수 있다. 어느 경우에든 상기 언어는 번역 또는 해석 언어(compiled or interpreted language)일 수 있다.

[0016] 바람직하기로는 본 발명의 프로그램은 저장 매체 또는 장치가 본원에 기술된 절차를 수행하도록 컴퓨터에 의해 판독되는 경우 컴퓨터를 구성 및 운영하기 위해, 범용 또는 전용의 프로그램가능 컴퓨터에 의해 판독가능한 저장매체 또는 장치에 저장될 수 있다. 본 발명의 시스템은 컴퓨터 프로그램을 사용하여 구성된 컴퓨터-판독가능 저장매체로서 이용되는 것으로 간주될 수 있으며 이 경우 그렇게 구성된 저장매체는 컴퓨터가 본원에 기술된 기능을 수행하도록 특정의 그리고 소정의 방식으로 작동하게 한다.

[0017] 본 발명에 있어서, '펄(Perl)'은 텍스트 파일로부터 필요한 정보를 추출하고 그 정보를 바탕으로 새로운 문서를

구성하는데 적합한 프로그래밍 언어이다.

[0018] 이하, 본 발명의 구체적인 내용을 실시예를 들어 상세히 설명한다.

[0019] **실시예 1: 인공 유전체 제조**

[0020] 염기서열이 이미 모두 알려진 20개의 미생물을 선정하여 각 미생물의 genomic DNA를 추출한 후 동일한 DNA 농도를 섞어서 인공 유전체(Mock community)를 제조하였다(표 1).

[표 1] Mock community의 구성

Strain	Genome size (bps)	Gene (# of gene copy)
Bacillus cerus ATCC 14579	5,427,083	16S (7)
Burkholderia vietamensis G4	8,391,070	16S (2), *nifH* (1)
Burkholderia xenovorans LB400	9,731,138	16S (2), *bphD* (1), *nifH* (1)
Chromobacterium violaceum ATCC 12472	4,751,080	16S (1)
Corynebacterium glutamin ATCC 13032	3,282,708	16S (9)
Desulfitobacterium hafniense DCB-2	5,279,134	16S (5), *nifH* (4)
Escherichia Coli K-12 sub W3110	4,646,332	16S (6)
Neisseria sicca ATCC 29256	2,830,772	16S (1), *nirK* (1)
Nostoc PCC 7120	7,211,789	16S (2), *nifH* (3)
Ochrobactrum anthropi ATCC 49188	5,205,777	16S (1), *nirK* (2)
Polaromonas naphthalenivorans CJ2	5,366,143	16S (1), *bphD* (1), *nifH* (1), *nirK* (1)
Pseudomonas pickettii PKO1	5,325,729	16S (1)
Pseudomonas putida F1	5,959,964	16S (3), *bphD* (1),
Rhodobacter sphaeroides KD 131	4,711,139	16S (1), *nifH* (3), *nirK* (1)
Rhodococcus sp. RHA1	9,702,737	16S (1), *bphD* (6),
Rhodospirillum rubrum ATCC 11170	4,406,557	16S (1), *nifH* (5)
Roseobacter denitrifican Och 114	4,331,234	16S (1)
Sphingobium yanoikuyae B1	5,915,246	16S (1)
Staphylococcus epidemidis ATCC 12228	2,564,615	16S (1)
Xanthomonas campestris ATCC 33913	5,076,188	16S (1)
Total : 20 strains		**16S (48), *bphD* (9), *nifH* (18), *nirK* (5)**

[0021]

[0022] 인공 유전체 내 미생물의 DNA는 Mobio사의 PowerMAX soil DNA extraction isolation kit를 사용하여 추출하고 PCR(Polymerase Chain Reaction)을 이용하여 16S rRNA 유전자 V1-V3 및 기능성 유전자 *bphD* 및 *nif*H를 증폭하였다. 모든 PCR 반응은 AccuPrimeTM Taq DNA Polymerase High Fidelity를 이용하여 수행하였다. PCR 산물은 QIAquick PCR purification kit를 이용하여 정제하였으며 Thermo Scientific사의 NanoDrop ND-1000을 이용하여 정량하였다. GS-FLX을 이용하여 PCR 산물의 티타늄 파이로시퀀싱을 수행하였다.

[0023] 비교예: 순수 배양

[0024] 인공 유전체 DNA의 파이로시퀀싱과 비교하기 위해 하기 4종의 미생물을 순수 배양하여 표준 시퀀스로 선정하였다.

[0025] *Staphylococcus epidermidis* ATCC 12228, *Roseobacter denitrificans* OCh 114, *Rhodococcus* sp. RHA1, *Polaromonas naphthalenivorans* CJ2

[0026] **실시예 2: 모호한 시퀀스의 발생 분석**

[0027] 본 실시예에서는 파이로시퀀싱 과정에서 발생할 수 있는 모호한 시퀀스(ambiguous bases; N's)가 특정 위치에 발생한다는 점에 집중하여 모호한 시퀀스의 발생 패턴을 분석하였다.

[표 2] 순수배양 유전자의 시퀀스 모티프(Sequence motif) 내 발생한 모호한 오류(N's)

Organism_gene	Dominant base at N (true base)	Sequence motif	N error rate per read
Staphylococcus_16S	C	CGATCACCCTNTCAGGTCGGC	0.0103
Roseobacter_16S	C	GAAACCCCGGNTAACTCCGTG	0.0144
Polaromonas_bphD	G	CGCTGGGGGTNTGTTCGAGCA	0.2328
Rhodococcus_bphD	T	GCGAACCTTCNCGG	0.0050

[0028]

[0029] 상기 [표 2]에서 밑줄친 글자는 모호한 시퀀스(N's) 앞의 호모폴리머를 나타낸다.

[0030] 모호한 시퀀스의 발생 분석 결과 모호한 시퀀스는 같은 염기서열이 반복되는 지점인 호모폴리머(homopolymer)의 1~2 베이스(bases) 하부(downstream)에 주로 발생한다는 것을 확인하였다. 즉, 상기 호모폴리머는 모호한 시퀀스의 상부(upstream)에 존재하였다. 또한 호모폴리머의 길이가 길어질수록 모호한 시퀀스의 발생이 증가하는 패턴을 확인하였다.

[0031] 본 실험예에 사용된 미생물의 염기서열정보 분석 결과와 기존의 알려진 미생물의 염기서열 분석 결과를 비교하여 모호한 시퀀스의 자리에 들어갈 실제 값을 분석한 결과, 모호한 시퀀스에 들어갈 실제 값은 앞서 발생한 호모 폴리머의 시퀀스 정보와 같다는 결과를 얻었다(표 2).

[0032] 모호한 시퀀스는 특정 위치에서 현저하게 증가하였는데 유전자 영역만이 아니라 미생물 종류에 따라서 모호한 시퀀스의 분포는 다양하였다(도 1).

[0033] 대부분의 N's은 호모폴리머 자리가 아니라 그 다음에 나오는 동일한 염기 자리에 발생하였다. 뉴클레오티드 반응물의 부족에 의한 비-동기화 연장의 증거는 N's의 염기 뒷부분에서 발견되지 않았다.

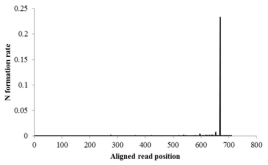
[0034] 길이가 긴 호모폴리머의 부산물은 DNA 합성에 관하여 유의적인 저해 없이 다음에 나오는 동일한 염기의 파이로포스페이트(pyrophosphate) 방출에 관한 선택적인 저해 효과를 가질 수 있었다.

[0035] 상기 실험 결과를 기반으로 모호한 시퀀스의 검측 및 수정이 가능한 'Perl'(펄) 기반의 프로그램을 발명하였으며 해당 프로그램을 이용하여 인공 유전체로부터 수득한 시퀀스를 보정해 본 결과 16S rRNA의 경우 약 86.54%, bphD 유전자의 경우 81.37%, nifH 유전자의 경우 81.55%의 모호한 시퀀스를 수정하였다(도 2).

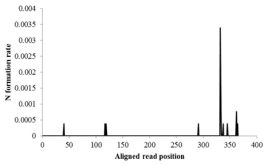
도면

도면1

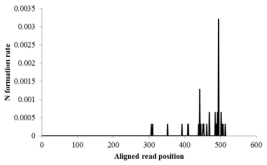
bphD region of *Polaromonas naphthalenivorans* CJ2



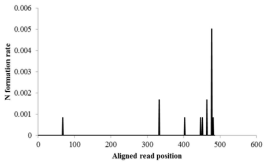
nifH region of *Polaromonas naphthalenivorans* CJ2



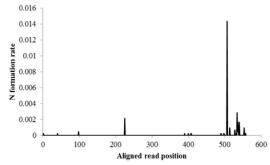
16S region of *Rhodococcus* sp. RHA1



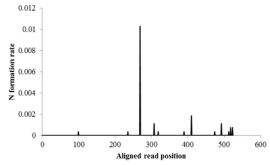
bphD region of *Rhodococcus* sp. RHA1



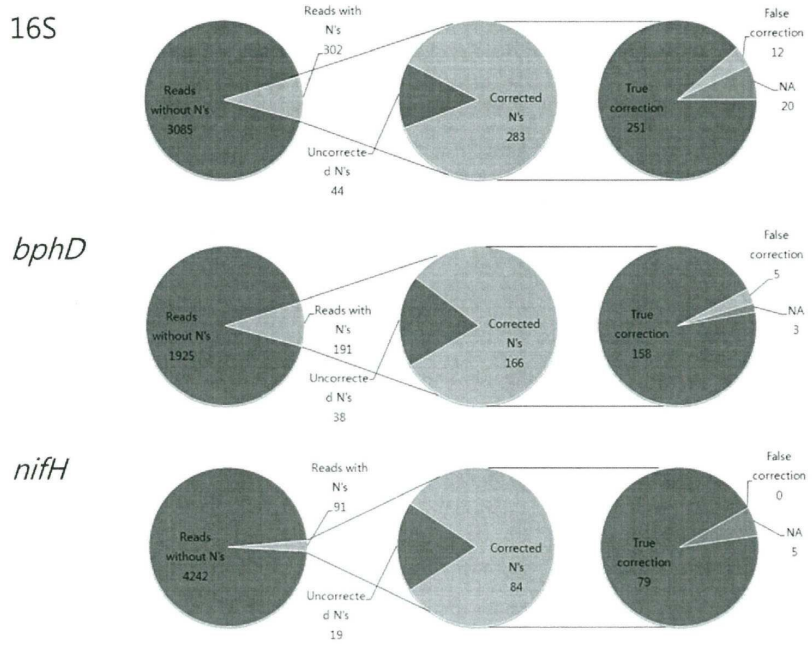
16S region of *Roseobacter denitrificans* OCh 114



16S region of *Staphylococcus epidermidis* ATCC 12228



도면2


16S
Reads without N's 3085
Reads with N's 302
Uncorrected N's 44
Corrected N's 283
True correction 251
False correction 12
NA 20
bphD
Reads without N's 1925
Reads with N's 191
Uncorrected N's 38
Corrected N's 166
True correction 158
False correction 5
NA 3
nifH
Reads without N's 4242
Reads with N's 91
Uncorrected N's 19
Corrected N's 84
True correction 79
False correction 0
NA 5

도면3

Supplementary Script S3 | Correction of reads with N's. This script was coded in Perl.

```
open(RF, "<input.txt");
open(WF, ">output.txt");
while($line=<RF>)
{
        if($line=~/^>/)
        {
                printf WF $line;
        }
        else
        {
                $line=~s/(AA[T]{1,4})N/$1A/g;
                $line=~s/(TT[G]{1,4})N/$1T/g;
                $line=~s/(GG[C]{1,4})N/$1G/g;
                $line=~s/(CC[A]{1,4})N/$1C/g;
                $line=~s/(AA[G]{1,4})N/$1A/g;
                $line=~s/(TT[C]{1,4})N/$1T/g;
                $line=~s/(GG[A]{1,4})N/$1G/g;
                $line=~s/(CC[T]{1,4})N/$1C/g;
                $line=~s/(AA[C]{1,4})N/$1A/g;
                $line=~s/(TT[A]{1,4})N/$1T/g;
```

도면4

```
$line=~s/(GG[T]{1,4})N/$1G/g;

$line=~s/(CC[G]{1,4})N/$1C/g;

$line=~s/(AA[G]{1,4}[T]{1,4})N/$1A/g;

$line=~s/(AA[C]{1,4}[G]{1,4})N/$1A/g;

$line=~s/(AA[C]{1,4}[T]{1,4})N/$1A/g;

$line=~s/(TT[C]{1,4}[G]{1,4})N/$1T/g;

$line=~s/(TT[A]{1,4}[C]{1,4})N/$1T/g;

$line=~s/(TT[A]{1,4}[G]{1,4})N/$1T/g;

$line=~s/(GG[A]{1,4}[C]{1,4})N/$1G/g;

$line=~s/(GG[T]{1,4}[A]{1,4})N/$1G/g;

$line=~s/(GG[T]{1,4}[C]{1,4})N/$1G/g;

$line=~s/(CC[T]{1,4}[A]{1,4})N/$1C/g;

$line=~s/(CC[G]{1,4}[T]{1,4})N/$1C/g;

$line=~s/(CC[G]{1,4}[A]{1,4})N/$1C/g;

$line=~s/(AA[C]{1,4}[G]{1,4}[T]{1,4})N/$1A/g;

$line=~s/(TT[A]{1,4}[C]{1,4}[G]{1,4})N/$1T/g;

$line=~s/(GG[T]{1,4}[A]{1,4}[C]{1,4})N/$1G/g;

$line=~s/(CC[G]{1,4}[T]{1,4}[A]{1,4})N/$1C/g;

printf WF1 $line;
```

도면5

```
close(RF);

close(WF);
```