BARM : 유전자 조립과 레퍼런스 얼라인먼트를 접목한 메타유전체 비닝 방법

(BARM : A Metagenome Binning Method using Genome Assembly and Reference Alignment)

여 윤 구 † 문 명 진 † 김 우 철 †† 박 상 현 †††
(Yunku Yeo) (Myungjin Moon) (Woocheol Kim) (Sanghyun Park)

요 약 메타유전체는 환경에서 직접 채취한 유전체 정보의 집합으로서, 이를 통해 연구실 환경에서 얻을 수 없는 다양한 유전체 정보를 얻을 수 있다. 메타유전체에는 수많은 생물체의 유전체가 뒤섞여 있기 때문에, 그 안에 존재하는 생물의 구성과 비율을 알아내는 것이 중요한 문제가 된다. 이러한 문제를 비닝이라고 하는데, 16S rRNA를 이용하는 것이 대표적인 비닝 방법이다. 16S rRNA를 이용하면 매우 정확한 결과를 얻을 수 있지만, 별도의 라이브러리를 구축해야 하기 때문에 시간과 비용이 많이 필요하다. 이 때문에, 컴퓨터 계산을 기반으로 하는 여러 가지 비닝 방법이 개발되고 있다. 본 논문은 레퍼런스 얼라인먼트 방식의 비닝 방법에 유전체 조립(genome assembly) 알고리즘을 융합하여 새로운 비닝 방법인 BARM을 개발하였다. 가상 변이 생성기를 이용하여 메타유전체 환경을 시뮬레이션하여 실험한 결과, 레퍼런스 데이터가 부족한 종의 비닝에 있어서 기존 비닝 방법보다 더 우수한 결과를 나타내었다.

키워드 : 메타유전체, 비닝, 유전체 조립, 레퍼런스 얼라인먼트

Abstract Metagenome is a large set of genomic information, collected directly from the environmental sample. We can acquire much information which cannot be obtained under laboratory conditions. Since the metagenome is a complicate mixture of numerous species, it is an important problem to infer the composition of species in metagenome – the binning problem. Binning with 16S rRNA is the most widely-used and accurate binning method. But it requires much time and high cost for the construction of additional biological library. To solve this problem, many computational binning methods such as Random Sequence Read(RSR) are developed. In this paper, we suggest a new binning approach BARM – a conjunction of reference alignment and genome assembly. We compare BARM with RSR using the synthetic genomic-variants generator, and BARM produced more superior result in genomic diversity of metagenome.

Key words : metagenome, binning, genome assembly, reference alignment

· 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2010-0008639)

† 학생회원 : 연세대학교 컴퓨터과학과
yyk@cs.yonsei.ac.kr
psiwind@cs.yonsei.ac.kr
†† 정 회 원 : 연세대학교 컴퓨터과학과
twelvepp@cs.yonsei.ac.kr
††† 종신회원 : 연세대학교 컴퓨터과학과 교수
sanghyun@cs.yonsei.ac.kr
(Corresponding author임)

논문접수 : 2010년 2월 3일
심사완료 : 2010년 11월 15일

정보과학회논문지: 데이타베이스 제38권 제2호(2011.4)

1. 서 론

메타유전체학(metagenomics)이란 유전체 연구의 한 방법론으로서, 바닷물, 생물체의 내장 속과 같은 실제 환경에서 여러 미생물의 유전체를 한꺼번에 채취하여 연구하는 방식이다. 전통적인 유전체 연구 방식은 연구실 환경에서 배양한 한 가지 생물의 유전체만을 단독으로 연구해 왔다. 그러나 자연에 존재하는 대부분의 미생물에 비해서 연구실 환경에서 배양할 수 있는 미생물의 종류는 매우 적기 때문에[1], 전통적인 방법만으로는 다양한 유전체 정보를 얻기 어렵다.

메타유전체학은 이러한 문제점을 해결하기 위하여, 미생물을 단독으로 배양하는 대신 환경에서 직접 채취한

다. 지금까지 연구된 메타유전체의 주요 환경으로는 광산의 침출수[2], 바닷물[3,4], 고래의 뼈[5] 등이 있으며, 환경에 따라 몇 종에서부터 수천 종의 미생물이 포함되어 있다. 이렇게 복잡하고 다양한 환경을 통해 연구실에서 얻을 수 없는 여러 미생물의 유전체 정보를 연구할 수 있으며, 생물종 간의 상호 작용, 특정 환경에서 서식하는 생물종의 공통적인 특징 등을 알아낼 수 있다. 메타유전체학 연구로 얻은 정보는 그 자체로도 가치가 높을 뿐 아니라, 이것을 이용해 기존에는 배양할 수 없는 미생물을 배양하는 데 필요한 지식을 얻을 수 있다.

메타유전체학에서는 유전체를 환경에서 직접 채취하기 때문에, 채취한 메타유전체에는 수많은 종의 유전체 정보가 뒤섞인 채로 존재하게 된다. 또한, 메타유전체 내에는 어떤 종은 풍부하게, 어떤 종은 매우 적은 수가 분포할 수도 있다. 뿐만 아니라, 같은 종 내의 생물일지라도 유전체 변이(polymorphism)가 크게 존재할 수 있다. 초기의 메타유전체 연구에서는 단일 유전체에 적용하던 유전체 조립(genome assembly) 방법을 그대로 적용하였으나, 메타유전체 구성의 복잡성, 변이와 같은 문제로 인하여 바닷물과 같이 복잡한 환경에서는 좋은 결과를 얻지 못했다.

이와 같은 문제점으로 인하여 메타유전체 내의 유전체를 연구하는 것은 매우 높은 복잡도를 갖게 된다. 또한 메타유전체 내에 존재하는 뒤섞인 유전체가 어떤 생물의 유전체인지를 알아내는 것 자체가 중요한 문제가 되었다. 메타유전체 내에 존재하는 생물의 종류와 분포를 알아내는 문제를 비닝(binning)이라고 하며, 이것은 메타유전체학에 있어서 중요한 목표 중 하나이다.

메타유전체의 비닝 방법 중 가장 많이 쓰이는 방법은 16S rRNA를 이용하는 것이다. 16S rRNA란 대부분의 미생물 유전체에 존재하는 짧은 염기 서열로서, 종에 따라 특이성을 보이면서도 계통상 유사한 종 사이에는 유사성을 가진다. 이것을 이용하면 대부분의 미생물을 계통학적으로 분류할 수 있다.

16S rRNA방법은 매우 높은 정확도를 갖는 반면, 별도의 생물학적 실험을 거쳐야 하기 때문에 많은 시간과 비용이 필요하다는 단점이 있다. 이를 극복하기 위하여 실제 실험을 필요로 하지 않는 다양한 알고리즘이 연구되었다. 이러한 방법들 중 대표적인 것으로는 올리고뉴클레오티드 빈도(oligo-nucleotide frequency)와 같은 유전체 표지(genome signature)를 이용하는 방법[6,7]과, BLAST와 같은 유사성 검색 툴을 이용하여 레퍼런스 유전체(reference genome)와의 유사성을 찾는 방법(레퍼런스 얼라인먼트, reference alignment) 등이 있다[8]. 이 방법들에 대한 자세한 내용은 2장에서 다룰 것이다.

그러나, 유전체 표지를 이용하는 방법은 메타유전체 연구의 기본 재료인 리드(read) 정도 크기의 유전체 조각에서는 유전체 특징이 잘 드러나지 않는다는 단점이 있으며, 레퍼런스 얼라인먼트 방법은 레퍼런스 유전체가 없는 종의 비닝 성능이 감소하는 단점이 있다. 이러한 문제점을 극복하기 위하여, 본 논문에서는 레퍼런스 얼라인먼트 방식의 비닝 방법과 유전체 조립 방식을 융합한 새로운 비닝 방법인 BARM을 개발하였다. 유전체 조립 알고리즘을 융합함으로써, 비닝에 사용하는 유전체 조각의 크기를 증가시켜 레퍼런스와의 유사성을 더 잘 탐색할 수 있다. 또한, 유전체 조립 알고리즘을 통하여 비슷한 리드를 한꺼번에 묶어서 비닝함으로써 메타유전체 내부의 유사성 정보까지 동시에 고려할 수 있다. 이를 통해 메타유전체 내의 변이와 유전적 다양성에 더 우수한 결과를 나타내었다.

본 논문의 구성은 다음과 같다. 2장에서는 여러 가지 비닝 방법에 관련된 기존 연구들을 다루며, 3장에서는 본 논문이 제안하는 방법에 대해 설명한다. 4장에서는 데이터 셋과 실험 과정에 대해 설명하며, 5장에서는 실험 결과를 분석한다.

2. 관련 연구

메타유전체의 비닝에 이용되는 대표적인 방법으로는 유전체 조립, 유전체 표지, 레퍼런스 얼라인먼트 등이 있다. 그 중 유전체 조립 방법은 단일 유전체에서 사용하던 조립 알고리즘을 메타유전체에 적용하는 것이다. 초기의 메타유전체 연구에서 이 방법을 적용해 보았지만, 유전체의 변이와 메타유전체 내 생물종의 커버리지 문제로 인해 유전체가 잘 조립되지 않았다.

유전체 표지를 이용하는 방법은 종 특이적(species-specific)인 속성을 찾아 비닝에 이용하는 것이 핵심이다. 올리고뉴클레오티드 빈도는 대표적인 유전체 표지 중 하나로서, 어떤 유전체 내에서 짧은 길이의 염기서열(oligo-nucleotide)의 상대적인 출현 빈도를 의미한다. 올리고뉴클레오티드의 길이를 n이라 할 때, n-뉴클레오티드는 총 4n개의 종류를 갖게 되며, 이 n-뉴클레오티드의 빈도는 곧 유전체의 특징을 표현하는 4n 차원의 벡터가 된다. David 등[9]의 연구에 따르면, 이러한 올리고뉴클레오티드의 빈도는 종 별 특이성을 반영하고 있으며, Teeling 등[10]은 같은 종에서 추출한 유전체 조각 사이에서는 이 빈도가 상대적으로 유사하게, 다른 종에서 추출한 조각 사이에서는 올리고뉴클레오티드의 빈도가 상대적으로 다르게 나타나는 것을 확인하였다. 올리고뉴클레오티드 빈도를 유전체 조각의 특징으로 사용하고 SOM(Self-Organizing Map)을 이용해 메타유전체 조각을 군집화(clustering)하려는 연구도 있었다[6,7].

그러나 이러한 방법들은 올리고뉴클레오티드의 빈도가 유전체 전체에 걸쳐서 고른 값을 가진다는 것을 기본으로 가정하고 있다. 그러나 논코딩 영역(noncoding region)과 같은 유전체의 세부 영역을 대상으로 했을 때에는 이러한 가정이 항상 성립하지 않는다[11]. 또한 메타유전체의 리드와 같이 작은 유전체 조각은 전체 유전체의 경향을 정확히 반영하기 어렵기 때문에 유전체 표지 방법을 사용하기 어렵다. Teeling 등[10]의 실험에서는 40Kbp 크기의 유전체 조각을 사용했으며, Abe 등[8]의 연구에서는 10Kbp 크기의 유전체 조각을 사용했다. 또한 우리는 본 논문의 사전 연구[12]에서 올리고뉴클레오티드 빈도를 추출하는 유전체 조각의 크기와 유전체의 특징 보존 정도와의 상관관계를 실험해 보았고, 그 결과 유의미한 유전체 특징을 확인하기 위해서는 7kbp 이상까지 유전체 조각의 크기를 증가시켜야 함을 확인했다. 일반적인 유전체 리드의 크기가 500~700bp라는 것을 감안할 때, 올리고뉴클레오티드 빈도를 이용하는 방법들을 메타유전체 리드에 적용하기 어렵다. Andrey. K 등[7]의 연구에서는 400bp 크기의 유전체 조각을 사용하기는 하였으나, 메타유전체 구성의 복잡도와 유전체의 특징에 따라 결과에 많은 영향을 받았다.

레퍼런스 얼라인먼트 방식은 데이터베이스에 있는 알려진 유전체 정보와 메타유전체의 유사성을 검색하여 비닝하는 방법이다. Manichanh 등[8]은 RSR(Random Sequence Read) 방법을 제안하였는데, 메타유전체의 연구를 위해 생성된 리드 라이브러리(read library)와 로컬 얼라인먼트(local alignment) 프로그램인 BLAST[13]를 이용하는 방식이다. 먼저 리드 라이브러리에서 임의로 고른 리드를 BLAST를 이용하여 이미 알려진 유전체 데이터베이스에서 유사성을 검색한다. 이후 검색 결과의 cutoff, e-value 등을 감안하여 메타유전체 내의 생물 구성을 추정하였다. 그 결과 16S rRNA 방법에 비해 정확성은 떨어지지만, 빠르고 저렴하게 유사한 생물 구성을 추정할 수 있었다. Monzoorul H 등[14]은 번역(translation)된 유전체 서열과 그것의 발현인 단백질 데이터베이스를 검색하는 BLASTx를 이용하였다.

레퍼런스 얼라인먼트 방식은 일반적으로 이미 존재하는 유전체 서열의 경우 매우 높은 정확도로 존재 여부를 알아낼 수 있다. 반면, 알려진 유전체 서열과 전혀 다른 종의 경우에는 그것이 어떤 생물인지 알아내기 어렵다. 즉, 레퍼런스 얼라인먼트 방법의 가장 큰 문제점은 데이터베이스에 있는 유전체 정보와 메타유전체의 유전체 정보가 다르거나, 혹은 데이터베이스에 유전체 정보가 존재하지 않을 때 비닝을 수행할 수 없다는 점이다.

3. 제안하는 방법

RSR방법은 레퍼런스 유전체(이미 알고 있는 유전체 정보)와 메타유전체 리드간의 1:1 유사성만을 기반으로 비닝을 수행하였다. 이 때문에, 그림 1과 같은 경우, 올바른 답을 제공하지 못할 수 있다. 그림 1(a)는 데이터베이스에 현재 종의 유전체 중 일부가 존재하지 않는 경우를 나타낸다. 이 경우에 데이터베이스와 리드간의 1:1 유사성만을 가지고 판단하면, 점선으로 표시한 리드는 유사한 생물종을 찾지 못하게 된다. 그림 1(b)는 유전체의 일부에 변이가 존재하는 경우를 나타낸다. 이 경우에는 유전체 서열의 많은 부분이 데이터베이스와 일치하지 않기 때문에 원래 종으로 비닝되지 않을 수 있다. 따라서 레퍼런스 얼라인먼트 방법은 이렇게 데이터베이스 내에 정확한 데이터가 존재하지 않을 경우, 메타유전체 내의 생물종 구성과 비율을 잘못 판단하게 된다.

이러한 문제를 해결하기 위하여, 본 논문의 비닝 방법은 레퍼런스 리드와 메타유전체 리드와의 1:1 유사성 뿐만 아니라, 메타유전체 리드끼리의 유사성까지 확장하여 고려한다. 메타유전체 리드끼리의 유사성을 빠르게 검색하기 위하여 유전체 조립 알고리즘을 적용하며, 조립 알고리즘을 통해 연결된 메타유전체 리드를 함께 비닝하는 것이 본 논문의 핵심 아이디어이다. 본 논문의 방법은 먼저 이미 알려진 종의 유전체 서열을 리드 크기로 잘게 조각낸 뒤, 이 조각들(이후 레퍼런스 리드라고 표

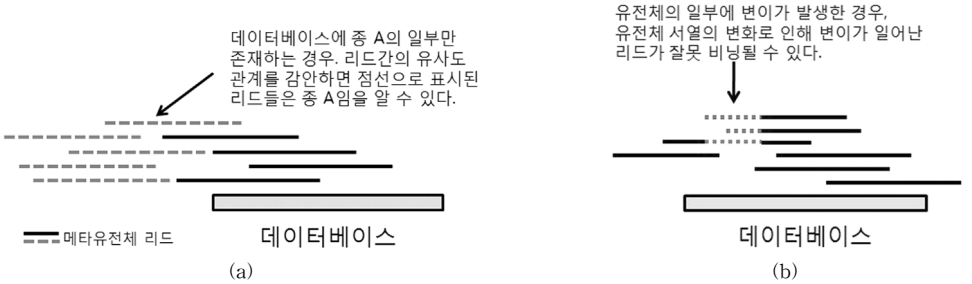


그림 1 데이터베이스와 메타유전체 리드간의 1:1 유사성을 기반으로 할 때의 문제점

기함)에 원래의 종을 알 수 있도록 표지를 부착한다. 이후 레퍼런스 리드를 메타유전체 리드와 섞어서 간략한 조립 작업을 수행한다. 이와 같이 유전체 조립 알고리즘을 적용하면 레퍼런스 리드와 메타유전체 리드가 조립될 뿐만 아니라, 메타유전체 리드와 메타유전체 리드 또한 조립된다. 이를 통해 레퍼런스 리드와의 유사성과 다른 메타유전체 리드와의 유사성을 동시에 고려할 수 있게 되고, 이런 정보를 이용하여 그림 1에서 표시된 리드들을 원래 종으로 분류할 수 있다.

본 논문에서 제안하는 방법은 다음과 같은 오버랩 탐색(overlap detection), 어셈블리 형식 비닝(assembly-like binning)의 2단계로 구성되어 있다.

3.1 오버랩 탐색

본래 유전체 조립 알고리즘은 조각난 여러 개의 유전체 서열을 유사성에 기반하여 하나로 연결하여 완결된 하나의 서열(consensus sequence)을 찾는 것을 목적으로 한다. 따라서 하나의 완결된 유전체 정보를 생성하기 위해, 수많은 염기 서열 간의 유사성을 모두 비교해야 한다. 이를 위해 사용되는 동적 프로그래밍 과정 때문에 유전체 조립 알고리즘이 많은 시간을 소요하게 된다. 그러나, 본 연구에서는 완결된 서열을 만드는 것이 목적이 아니라, 리드간의 유사성을 확인하는 것으로 충분하다. 따라서 유전체 조립 알고리즘의 전체를 적용할 필요가 없다. 또, 같은 종이라 할지라도, 서로 다른 개체에서 생성된 리드일 경우 유전체 변이로 인해 염기 서열 간에 차이가 있을 수 있다. 이런 두 리드 사이의 완전한 얼라인먼트를 계산할 경우, 유전체 변이로 인하여 리드들이 잘 조립되지 않고 작은 조각들로 남게 되고, 유전체 조립 알고리즘의 효율이 하락하게 된다.

이러한 문제를 동시에 해결하기 위하여, 본 논문에서는 일반적인 유전체 조립 알고리즘의 일부만을 변형·적용하였다. 본 논문에서 사용한 알고리즘은 arachne 어셈블러[15]의 오버랩 탐색 알고리즘을 수정한 형태이다. Arachne의 오버랩 탐색 알고리즘은 fasta 알고리즘[16]과 유사한 것으로, 양쪽 유전체 서열에 존재하는 k 길이의 워드(k-mer)를 기본 단위로 이용하여 유전체 서열 간의 유사성을 탐색하는 알고리즘이다. 유전체 서열로부터 k-mer를 추출할 때에는 유전체가 갖는 정보가 손실되지 않도록, 그림 2(a)와 같이 1bp 단위의 슬라이딩 윈도우(sliding window)를 이용하여 추출한다.

본 논문에서 사용한 k값은 32이다. k 값을 크게 설정하면 한 차례의 비교로 더 큰 길이의 염기 서열을 비교할 수 있게 되지만, 이 경우 존재할 수 있는 모든 k-mer의 종류는 432이 되고, 모든 경우의 수를 저장하려면 엄청난 크기의 메모리를 필요로 하게 된다. 그러나 실제로는 유전체 서열의 전체 크기조차도 432보다 훨씬 작으며, 염기 서열 간의 유사성으로 인해 많은 수의 중복이 존재한다. 때문에 전체 k-mer가 아니라 실제로 비교할 서열에 존재하는 k-mer만을 저장하면 메모리 사용량이 크게 감소하게 된다.

여기에 본 논문에서는 k-mer의 표현 방식을 4진수 기반으로 부호화함으로써 필요 메모리를 훨씬 감소시켰다. 본 논문에서는 k-mer를 k개의 문자열 그대로 저장하는 대신, 그림 2(b)와 같이 A~T의 염기 서열을 각각 0(4)~3(4)의 4진수로 치환하였다. 이것은 1개의 문자(8bit)가 2bit로 표현될 수 있음을 의미한다. 이를 통해 k-mer의 저장에 필요한 메모리를 1/4로 감소시킬 수 있다. 이렇게 k-mer의 크기가 작아지면 추후 k-mer를 비교하여 얼라인먼트를 찾을 때의 속도까지 향상시킬 수 있다. 본 논문에서 k값으로 32를 사용한 것은 long int(64bit) 변수 하나에 k-mer의 데이터를 저장할 수 있기 때문이다. 두 개의 k-mer가 같은 값인지 확인하는 과정을 별도의 함수 호출 없이 64비트 정수 변수의 == 연산으로 수행할 수 있었다. k-mer를 나타내는 데이터 구조는 {k-mer가 속한 리드의 ID, k-mer를 나타내는 64bit integer, 리드에서의 위치, 방향}로 구성되어 있다. 유전체 서열을 정수값의 k-mer로 변형할 때는, 그것을 5′ 방향(정방향)으로 읽었을 때의 값과 3′ 방향(역방향)로 읽었을 때의 값을 함께 계산한다. 이후 5′과 3′ integer 값 중 더 큰 값을 대표값으로 저장하고, 5′/3′ 여부를 저장한다. 이를 통해, 서로 역방향에 있는 리드간의 k-mer 비교도 64비트 정수값의 비교만으로 수행할 수 있다.

오버랩 탐색 알고리즘은 메모리를 절약하기 위해 arachne 어셈블러와 동일하게 100개의 연산 과정으로

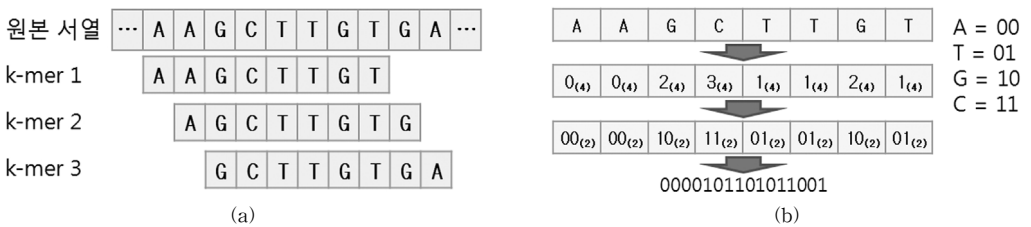


그림 2 원본 서열에서 k-mer를 추출하는 과정

나뉘어 진행된다. 첫 번째 연산 과정은 AA로 시작해서 AA로 끝나는 k-mer와 그것의 역상보 서열(reverse complement)을 검색하며, 두 번째 연산 과정은 AA...AT인 k-mer와 그것의 역상보 서열을 검색한다. 역상보 서열까지 동시에 검색하는 이유는, 유전체 조각 사이의 유사성을 찾을 때에 두 조각이 서로 반대 방향으로 배열되었을 때까지 고려해야 하기 때문이다. AA...AA를 검색할 때 TT...TT를 함께 검색하기 때문에 연산 과정을 1회 줄일 수 있다. 단, AT...AT와 같이 역상보 서열이 원래와 똑같은 경우는 위와 같은 pass 감소가 생기지 않는다. 이와 같이 연산 과정을 줄이고 나면 총 100번의 연산 과정으로 존재하는 모든 k-mer를 처리할 수 있다.

연산 과정 하나의 개괄적인 알고리즘은 그림 3과 같다. 먼저 모든 유전체 서열 조각에서 현재 연산 단계에 해당하는 k-mer를 추출하여 배열 L을 작성한다(그림 3의 line 1. prefix와 postfix는 현재 연산 단계를 나타내는 뉴클레오티드를 나타낸다). 배열 L을 k-mer값을 기반으로 정렬하면, 같은 k-mer끼리 근접하여 모이게 된다(line 2). 두 리드가 같은 k-mer를 갖고 있다는 것은 k 길이만큼의 오버랩을 갖는다는 것과 같다. 따라서 L에서 같은 k-mer값을 갖는 부분을 부분 배열로 분리하여(line 7), 그 부분 배열 내에 속한 리드 간에서만 얼라인먼트를 찾음으로써 대상 구간을 단축할 수 있다. 여기까지의 오버랩 탐색 과정은 arachne 어셈블러의 알고리즘과 동일하게 진행된다. 그 후 부분 배열 내에 존재하는 모든 k-mer 조합에 대하여(line 8), 각각의 k-mer를 포함하고 있는 두 리드를 찾은 뒤, 두 리드 사이의 얼라인먼트를 계산한다(line 9~18). 이 때 이전에 계산이 완료된 리드의 조합은 다시 계산하지 않도록 한다.

두 리드 간의 얼라인먼트를 계산할 때에는 먼저 양쪽의 리드에서 k-mer를 추출한다. 이 때에는 현재 연산 과정의 k-mer와는 상관없이 존재하는 모든 k-mer를 추출한다. 이후, 양쪽 리드에서 공통으로 존재하는 k-mer의 목록을 작성한다(line 12). 이 때 공통 k-mer hit마다 {양쪽 리드에서의 위치, 양쪽 k-mer의 5'/3' 여부, match의 점수}를 기록한다. 마지막으로 공통 k-mer의 위치와 방향을 이용하여, 같은 방향을 가지면서 연속된 공통 k-mer 를 찾아 하나로 합쳐 나간다(line 13). 이 작업은 두 단계로 이루어진다. 첫 번째 단계는 연속적이며 중복된 영역을 갖는 k-mer를 찾아 합치는 단계이다. k-mer를 1bp 단위로 슬라이딩하면서 추출하기 때문에, 긴 길이의 얼라인먼트가 여러 개의 오버랩된 공통 k-mer 로 분할되어 나타난다. 따라서 이 첫 번째 단계는 이러한 긴 길이의 얼라인먼트를 찾아내는 단계에 해당한다. 두 번째 단계는 연속적이지 않은 공통 k-mer 사이의 모든 조합을 탐색하면서 갭(gap)에 페널티를 부여하며 합치는 단계이다.

일반적으로 두 염기 서열 간에 갭 페널티(gap penalty)를 적용할 때는, 시작 페널티(gap open penalty)와 확장 페널티(gap extension penalty)를 따로 적용하지만, 본 논문에서는 메타유전체에 있는 유전체 변이를 감안하기

```
Algorithm : 오버랩 탐색
Input     : 유전체 리드 집합 R
Output    : 오버랩을 갖는 리드 쌍과 점수의 목록 {(a, b), score}

01  L = extract_current_k_mer( R, prefix, postfix ); // 모든 리드에서 현재 단계에 해당하는 k-mer 추출
02  sort( L );
03  initialize overlap_list, visit_list;
04
05  while ( L is not empty )
06      k_current = first element of L;
07      L_sub = find_sub_array( L, k_current );  // k_currnet와 같은 값을 갖는 소구간을 선택
08      foreach ( k1, k2 ) in L_sub
09          R1 = origin_read( k1 );
10          R2 = origin_read( k2 );
11          if ( visit_list is not contain ( R1, R2 ) ) {
12              k_mer_list = find_common_k_mer( R1, R2 );
13              score = merge_k_mer( k_mer_list );
14              if ( score ≥ threshold T )
15                  add {(R1, R2), score} into ovarlap_list;
16              end
17              add (R1, R2) into visit_list;
18          end
19      end
20      remove L_sub from L;
21  end
22
23  return overlap_list;
```

그림 3 오버랩 탐색 알고리즘의 수도 코드

위하여 모든 갭에 동일한 페널티를 부여하였다(linear gap penalty). 위와 같은 기준으로 스코어를 계산한 결과가 임계값 T 이상이면 두 유전체 서열 간에 잠재적인 유사성이 있는 것으로 간주하였다. 이와 같은 방법을 이용하면 약간의 유전체 변이에도 더 유연하게 적응할 수 있으며, 유전체 조립 알고리즘을 모두 적용하는 것보다 빠르게 서열 간의 유사성을 검색할 수 있다.

RSR 방법에 사용한 BLAST와 본 논문의 시간복잡도를 비교해 보면 다음과 같다. BLAST의 시간복잡도는 기본적으로 O(N×m×n)이며, 레퍼런스 데이터에 미리 인덱스를 만들어 놓음으로써 시간복잡도를 O(N×M×n)으로 줄일 수 있다(N = 리드의 개수, n = 리드의 평균 길이, M = 레퍼런스의 개수, m = 레퍼런스의 길이). 여기서 M이 상대적으로 작은 값이기 때문에, BLAST의 시간복잡도는 O(N×n)에 근접하게 된다.

본 논문의 알고리즘은 전체 리드의 집합을 4k개의 소그룹으로 분할하고 같은 소그룹 내의 리드 사이에서만 비교를 수행하기 때문에, 리드간의 비교 회수가 $O(n^2/4^k)$로 감소한다. 또한 두 리드 사이의 얼라인먼트를 계산할 때에도 양쪽 리드에 공통으로 존재하는 k-mer에 대해서만 확장을 시도하기 때문에 양쪽 염기 서열 자체에 대한 얼라인먼트를 구하는 BLAST보다 계산량이 훨씬 적다. 동일한 조건의 컴퓨터에서 RSR 방법과 본 논문의 비닝 방법을 실행해 본 결과, 표 1과 같이 약 1/5의 시간만이 소요되었다. 오버랩 탐색의 결과물로 {메타유전체 리드, 레퍼런스 리드} 집합 내에서의 오버랩 목록이 생성되며, 이를 이용해 다음 단계의 비닝을 진행한다.

표 1 BLAST와 BARM의 수행시간 비교

구분	BLAST(blastall)	BARM
수행시간	약 35시간	약 7시간

3.2 어셈블리 형식 비닝(Assembly-like binning)

오버랩 탐색 알고리즘이 완료된 후에는 그림 4의 알고리즘과 같이 각 리드의 비닝을 수행한다. 먼저 아직 비닝 결과가 없는 리드를 선택한 후, 선택한 리드와 직접 또는 간접 연결을 갖는 모든 리드를 탐색한다(line 3~5). 직접 연결이란 어떤 리드와 리드가 직접 오버랩을 갖는 것을 의미하며, 간접 연결이란 직접 연결은 아니지만 다른 어떤 리드들을 중간에 거쳐서 연결된 것을 의미한다. 이후 연결된 모든 리드들을 탐색하면서, 리드와 연결된 레퍼런스 유전체의 목록을 작성하고 점수를 합산한다(line 7~13). 이 때 탐색한 리드들의 집합을 컨티그(contig)라고 부르기로 한다. 리드의 메이트 페어(mate pair)는 같은 클론(clone)에서 생성된 리드이기 때문에 같은 컨티그에 포함시킨다. 탐색이 완료되면 컨

```
Algorithm : 어셈블리 형식 비닝
Input     : 메타유전체 리드 집합 S
Output    : 각 리드의 비닝 결과. { ( R, species ) }
01    initialize binning_result;
02    while ( S is not empty )
03        R0 = first element of S;
04        initialize contig, ref_list;
05        contig = search_connected_read( R0 );
06
07        foreach R in contig
08            if ( R has one or more connections to a reference read K )
09                SK = species of K;
10                add SK into ref_list;
11                score( SK ) += score( R, K );
12            end
13        end
14
15        contig_species = find_best_score_species( ref_list );
16        contig_score = score( contig_species );
17
18        foreach R in contig
19            if ( R has one or more connections to a reference read K )
20                if ( score( R, K ) > contig_score )
21                    add ( R, species of K ) into binning_result;
22                else
23                    add ( R, contig_species ) into binning_result;
24                end
25            else
26                add ( R, contig_species ) into binning_result;
27            end
28        end
29
30        foreach R in contig
31            remove R from S;
32        end
33    end
34
35    return binning_result;
```

그림 4 어셈블리 형식 비닝 단계의 수도 코드

티그에 연결된 레퍼런스 유전체 중 점수가 가장 높은 것을 컨티그의 대표 종으로 결정한다(line 15~16). 이후 현재 컨티그에 속한 모든 리드를 순회하면서, 각 리드의 종을 결정한다(line 18~28). 각 리드를 비닝할 때는 현재 리드와 직접 연결된 레퍼런스 리드중에 가장 높은 점수를 갖는 연결 1개와, 컨티그 내의 메타유전체 리드 중에서 현재 리드와 가장 높은 점수를 갖는 직접 연결 1개를 비교한다. 이것은 레퍼런스와의 연결 강도(score)와 주변 메타유전체 리드와의 연결 강도를 비교하는 작업이다. 메타유전체 리드와는 약한 연결을 갖고 있으면서 컨티그의 대표 유전체가 아닌 다른 레퍼런스 리드와 더 강한 연결을 갖고 있다면, 그것을 유전체 조립 과정의 오류로 판단하여 제외시키는 과정이다.

이와 같은 비닝 방법은 유전체 조립 알고리즘처럼 메타유전체의 리드들을 리드간의 유사성을 기반으로 여러 개의 집합으로 묶은 뒤, 각각의 집합별로 비닝을 하는 방식이다. 이로서 어떤 리드가 레퍼런스 유전체와 명시적인 유사성이 없다 하더라도, 주변 리드를 이용한 간접적인 유사성을 이용하여 비닝을 수행할 수 있다.

4. 데이터 셋 및 실험

4.1 데이터 셋

본 논문에서 사용한 데이터 셋은 103개 종의 유전체 정보를 포함하고 있다(부록 1 참조). 실험에 사용한 유전체는 K. Mavromatis 등[17]이 가상의 메타유전체 환

경 조성에 사용한 생물 종의 목록을 참고하여 선정하였다. 유전체 정보는 Genbank[18]에서 다운로드하였으며, 100종의 Bacteria 군과 3종의 Archea 군을 포함하고 있다. 전체 데이터셋 중에서 20개의 종을 가상의 메타유전체를 구성할 종으로 선택하였다(메타유전체로 선정된 종은 부록 1에 음영으로 처리되어 있다.). 그 중 8개의 종을 알려지지 않은 미생물로(이후 "미지 유전체"로 표기함) 선정하였다. 이 8개의 미지 유전체 중, *Syntrophobacter fumaroxidans*의 경우, 데이터베이스 내에 같은 목(order) 계통의 생물이 존재하지 않도록 하였다. 또, *Methanococcoides burtonii*는 데이터베이스 내에 같은 문(phylum) 계통의 생물이 3종류로 매우 적도록 구성하였다. 이것은 실제 메타유전체 연구에 있어서 참고할 레퍼런스 데이터가 많지 않은 상황을 구성하기 위함이다. 전체 103종의 데이터셋 중 알려지지 않은 8개 종을 제외한 95개 종은 알려진 유전체 서열로서, 데이터베이스를 구성하게 된다.

4.2 실험방법

K. Mavromatis 등의 실험에서는 가상의 메타유전체를 만들었지만, 자연 환경에 존재할 수 있는 다양한 변이가 반영되지 않았다. 본 논문에서는 실제 환경에 보다 근접한 가상 메타유전체를 만들기 위해 유전체 변이를 데이터에 반영하였다. 먼저 가상 변이 생성기[19]를 이용, NCBI에서 다운받은 완전한 유전체 정보에 임의의 유전체 변이를 적용하여, 각 종마다 100개의 개체(individual)를 생성하였다. 가상 유전 변이 생성기는 기존에 밝혀진 염기 서열을 바탕으로 가상의 유전 변이를 추가하여 새로운 염기 서열을 생성하는 툴이다. 가상 유전 변이 생성기로 생성할 수 있는 변이로는 단일 염기 변이(SNP), 1kbps 미만의 삽입(insertion), 1kbps 미만의 삭제(deletion), 1kbps 이상의 삽입인 CNV 증가, 1kbps 이상의 삭제인 CNV 감소, 역위(inversion)가 있다. 가상 유전 변이 생성기에 사용되는 매개 변수로는 발생 빈도, 길이, 복제 수가 있다. 발생 빈도는 특정 유전 변이가 일어날 확률을 의미하며, 길이는 유전 변이가 일어날 구간의 길이를 뜻하며, 복제 수는 삽입과 CNV 증가의 반복 횟수를 의미한다. 미생물의 개체별 유전 변이에 대한 연구는 많이 이루어지지 않았기에, 발생 빈도는 유전 변이가 전체 염기 서열 영역의 0~50%가 되도록 임의로 조절하였으며, 길이와 복제 수에 대한 범위는 사람의 database of genomic variants[20]와 dbSNP[21]를 참고하여 설정하였다.

변이가 적용된 가상의 개체를 생성한 후에는 메타유전체에 해당하는 미생물의 구성 비율을 결정하였다. 메타유전체는 환경에서 직접 채취하기 때문에, 환경에 풍부하게 서식하는 생물의 경우 메타유전체 내에서 차지하는 비율 역시 크고, 환경에서 적게 서식하는 생물은 비율이 작다. 미지 유전체의 구성 비율을 결정할 때에는 적은 구성 비율부터 많은 구성 비율까지 고르게 할당하였다.

다음으로, 앞에서 만든 100개의 가상 개체 중에서 메타유전체의 구성 비율만큼의 개체를 랜덤하게 선정하여 메타유전체 리드를 생성하였다. 이 때, 실제 라이브러리 구축 과정과 유사하도록 일정한 길이의 클론(clone) 단위로 유전체를 읽어들였다. 그 후, 클론의 양 끝에서 약 6~800bp의 시퀀스를 읽어들여 리드를 생성하였다. 클론의 길이는 플라스미드(plasmid) 클론을 가정하여 평균 2.7kbp의 길이와 400bp의 표준편차를 갖는 정규분포를 따르도록 하였다. 또한 라이브러리 구축 과정에서의 불확정성을 반영하기 위하여, 클론과 클론 사이에는 약 0~1.5kbp까지 길이를 무작위로 건너뛰도록 하였다. 단, 시퀀싱 과정에서는 에러 없이 정확한 염기 서열을 읽어들이는 것으로 가정하였다. 이미 가상 변이 생성기를 통해 염기 서열에 변이를 삽입하였기 때문이다.

다음으로 비닝을 위해 레퍼런스 데이터셋으로부터 레퍼런스 리드를 추출하였다. 레퍼런스 유전체는 모든 종에서 1번 씩만 생성하였으며, 메이트 페어와 갭이 없이 1kbp 길이로 생성하였다. 각 레퍼런스 리드와 메타유전체 리드에는 원래 종을 알 수 있도록 표지를 부착하였다. 이후 메타유전체 리드와 레퍼런스 리드를 섞고, 한꺼번에 BARM의 방법을 적용하였으며, BLAST와 레퍼런스 데이터베이스를 이용해 레퍼런스 얼라인먼트 방법을 적용하였다.

5. 결과 분석

그림 5는 원본 메타유전체의 구성 비율과 비닝 결과로 추정한 구성 비율을 그래프로 나타낸 것이다. 구성 비율은 목(order) 수준을 기준으로 구성하였다. 전반적으로 원본 메타유전체의 구성과 흡사한 결과를 나타낸 것을 확인할 수 있다.

그림 5에서 네모로 표시한 *Methanosarcinales*와 *Pseudomonadales*의 경우, RSR에서는 구성 비율이 실제에 비해 매우 작게 나타났다(메타유전체에서는 각각 5.343%, 7.478%, RSR의 결과에서는 각각 0.036%, 0.428%). 이 2종의 계통은 가상 메타유전체 내에서 미지 유전체로만 존재하면서, 데이터베이스 내에 같은 계통의 유전체가 소수만 존재하는 경우였다. 그에 비해, 메타유전체 내에 레퍼런스 생물이 존재하는 계통은 원래 비율과 흡사한 결과를 나타냈다. 이것은 데이터베이스 내에 미지 유전체과 가까운 계통의 생물이 적거나, 또는 가까운 계통의 생물이 존재하더라도 유전체 서열이 일부만 유사할 경우, 충분한 얼라인먼트를 찾아내지 못했기 때문이다.

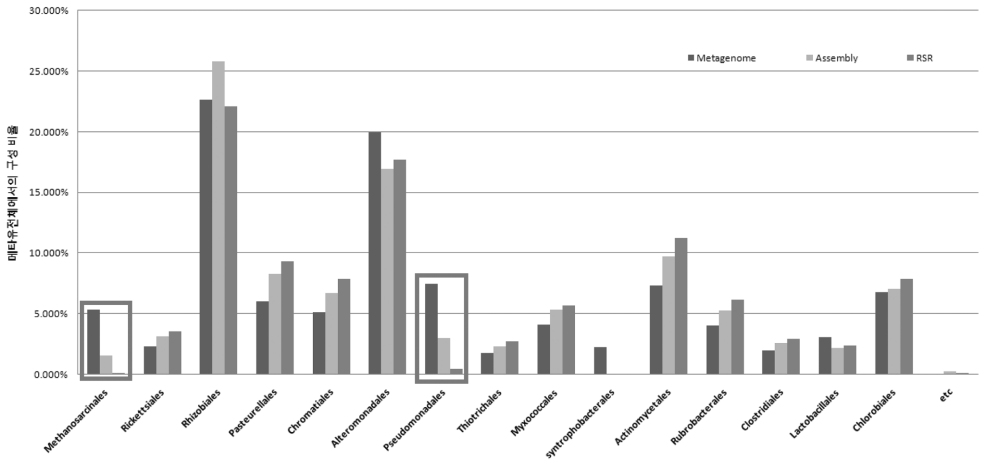


그림 5 메타유전체와 비닝 결과에서의 목(order) 기준 구성 비율

RSR 방법은 리드와 데이터베이스 간의 1:1 유사성만을 기반으로 하기 때문에, 이와 같은 경우 아주 적은 수의 리드밖에 비닝할 수 없다. 반면, 유전체 조립 알고리즘을 이용한 BARM의 경우 원본 메타유전체의 구성 비율에 비해서는 낮지만, 해당 종의 존재 여부를 파악할 수 있을 정도의 구성 비율이 나타났다(각각 1.546%, 3.001%. 표 2 참조). 유전체 조립 방법을 적용할 경우, 데이터베이스와 적은 수의 얼라인먼트가 발생하더라도 리드간의 연결 관계를 이용하여 많은 수의 리드를 비닝할 수 있다. 즉, 데이터베이스 내의 데이터가 모자란 생물 계통이 메타유전체에 존재하더라도 RSR 방법에 비해 더 높은 확률로 유사한 계통의 생물이 존재함을 알 수 있다.

다만, 양쪽 방법 모두 레퍼런스 얼라인먼트 방식의 비닝이기 때문에, 데이터베이스에 존재하지 않는 계통의 경우에는 그 존재 여부를 알아내지 못했다(*Syntrophobacterales*). 이 경우, 양쪽 방법 모두 보다 상위 계통의 다른 종으로 비닝되었다.

표 3은 정답을 알고 있다고 가정할 때, 메타유전체 실험셋에 대한 RSR 방법과 BARM 방법의 정답률을 정리한 결과이다. 메타유전체 내의 레퍼런스 종을 정확히 판정하는 것과 미지 유전체를 분류하는 부분에서는 전반적으로 RSR 방식과 대등한 결과가 나타났다. RSR방식은 약 10만 개의 리드에 대한 결과를 판정하지 못했다. 이것은 해당 리드의 얼라인먼트가 identity, e-value의 임계값을 통과하지 못했기 때문이다(기준이 되는 임계값은 Manichanh 등[8]이 사용한 값을 적용하였다). BARM 방법은 모든 리드에 대한 결과를 판정하였지만, 대신 레퍼런스 종을 잘못 판정하는 경우와 미지 유전체를 레퍼런스종으로 판정하는 경우가 RSR에 비해 많았다. 그러나 BARM 방법에서 종을 잘못 판정한 경우를 분석해 보면(표 4), 종이 잘못 판정된 경우라 할지라도 87% 이상의 리드가 같은 속(genus) 내의 생물종으로 비닝되었으며, 95% 이상의 리드가 과(family) 수준까지 동일한 계통으로 비닝된 것을 알 수 있다. 그렇기 때문에 BARM 방법의 결과를 가지고 메타유전체 내의 구성을 파악하는 데 지장이 없었다.

표 2 미지 유전체의 비닝결과 정리

목(Order)	DB 내 종 수	메타유전체의 원래 비율	BARM에서의 비율	RSR에서의 비율	비 고
Methanosarcinales	2	5.343%	1.546%	0.036%	메타유전체에 미지 유전체로만 존재
Pseudomonadales	5	7.487%	3.001%	0.428%	메타유전체에 미지 유전체로만 존재
Syntrophobacterales	0	2.226%	0%	0%	메타유전체에 미지 유전체로만 존재
Rhizobiales	7	22.648%	25.828%	22.077%	메타유전체에 레퍼런스가 존재
Alteromonadales	9	19.946%	16.964%	17.710%	메타유전체에 레퍼런스가 존재
Lactobacillales	8	3.047%	2.138%	2.364%	메타유전체에 레퍼런스가 존재
Chlorobiales	5	6.742%	7.018%	7.879%	메타유전체에 레퍼런스가 존재

표 3 메타유전체 실험셋에 대한 RSR 방법과 BARM을 이용한 방법의 결과 대조표

구 분	R S R		B A R M	
	개 수	비율(%)	개 수	비율(%)
전체 리드 개수	626,832	100.00	626,832	100.00
레퍼런스 종을 바르게 판정	322,625	51.47	337,457	53.84
미지 유전체를 없는 종으로 판정	142,667	22.76	156,474	24.96
레퍼런스 종을 없는 것으로 판정	68	0.01	8,076	1.29
레퍼런스 종을 다른 레퍼런스 종으로 판정	36	0.01	979	0.16
미지 유전체를 레퍼런스 종으로 판정	58,562	9.34	123,846	19.76
결과 판정 불가	102,874	16.41	0	0.00

표 4 BARM의 결과에서 종을 잘못 판정한 경우에 대한 유사성 분석표

C a s e	개 수	비 율
전체 오류 (레퍼런스를 잘못 판정+미지 유전체→레퍼런스 판정)	124,825	100.00
완전히 다른 종으로 비닝	3	0.002
계(superkingdom)까지 같은 종으로 비닝	224	0.179
문(phylum)까지 같은 종으로 비닝	505	0.405
강(class)까지 같은 종으로 비닝	2,087	1.672
목(order)까지 같은 종으로 비닝	2,012	1.612
과(family)까지 같은 종으로 비닝	10,359	8.299
속(genus)까지 같은 종으로 비닝	109,635	87.831

6. 결론 및 향후 연구

환경에서 직접 채취한 메타유전체는 다양한 유전적 정보를 포함하고 있지만, 그만큼 복잡도도 매우 높다. 그 때문에 메타유전체의 구성 비율을 파악하는 비닝 문제가 매우 중요하며, 다양한 방법으로 연구되고 있다. 레퍼런스 얼라인먼트 방식의 비닝은 이미 연구되어 데이터가 존재하는 생물 계통은 비교적 정확하게 찾을 수 있는 반면, 아직 연구되지 않은 생물 계통은 잘 찾을 수 없다. 본 논문은 레퍼런스 얼라인먼트 방법의 비닝에 유전체 조립 알고리즘을 융합하여, 레퍼런스 유전체와 리드와의 1:1 관계 뿐만 아니라 주변 리드와의 연결 관계까지 한꺼번에 고려할 수 있는 비닝 방법을 개발하였다. 새로운 비닝 방법은 데이터베이스의 내의 계통 정보가 부족한 경우에도 기존의 레퍼런스 얼라인먼트 방식의 비닝 방법보다 우수한 결과를 얻을 수 있었다.

BARM의 수행 과정에서 생성하는 컨티그는 유사한 종에서 추출된 것으로 추정되는 리드들의 집합이다. 이 컨티그에서 올리고뉴클레오티드 빈도와 같은 유전체 표지를 추출한다면, 짧은 길이의 리드보다 전체 서열의 특징을 더 뚜렷하게 보존하고 있을 가능성이 있다. 이를 통해 레퍼런스가 없어 비닝하지 못한 컨티그들을 군집화하거나, 비닝한 컨티그들을 유전체 표지 비교를 통해 다시 한 번 검증할 수 있을 것이다.

참 고 문 헌

[1] Norman R. Pace, "A Molecular View of Microbial Diversity and the Biosphere," *Science*, vol.276, no.5313, pp.734-740, 1997.

[2] Gene W. Tyson et al., "Community structure and metabolism through reconstruction of microbial genome from environment," *Nature*, 428, pp.37-43, 2004.

[3] Douglas B. Rusch et al., "The Sorcerer II Global Ocean Sampling Expedition : Northwest Atlantic through Eastern Tropical Pacific," *PLoS Biology*, vol.5, issue 3, pp.398-431, 2007.

[4] J. Craig Venter et al., "Environmental Genome Shotgun Sequencing of the Sargasso Sea," Science, vol.304, no.5667, pp.66-74, 2004.

[5] Susannah G. Tringe et al., "Comparative Metagenomics of Microbial Communities," *Science*, vol.308, no.5721, pp.554-557. 2005.

[6] Takashi Abe et al., "Informatics for Unveiling Hidden Genome Signatures," *Genome Research*, vol.13, pp.693-702, 2003.

[7] Andrey Kislyuk et al., "Unsupervised statistical clustering of environmental shotgun sequences," *BMC Bioinformatics*, vol.10, no.316, 2009.

[8] Chaysavanh Manichanh et al., "A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library," *Nucleic Acids Research*, vol.36, no.16 pp.5180-5188, 2008.

[9] David T. Pride et al., "Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases," *Genome Research*, vol.13, pp.145-158, 2003.

[10] Hanno Teeling et al., "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol.6, issue 9, pp.938-947, 2004.

[11] John Bohlin et al., "Investigations of Oligonucleotide Usage Variance Within and Between Prokaryotes," *PLoS Computational Biology*, vol.4, issue 4, 2008.

[12] Y. K. Yeo et al., "A Comparison of genome signature with various length genome fragments based on oligonucleotide frequency," *Proc. of the KIISE Korea Computer Congress 2009*, vol.36,

no.1(A), pp.58-59, 2009.
[13] Stephen F. Altschul et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids*, vol.25, no.17, pp.3389-3402, 1997.
[14] Monzoorul Haque M. et al., "SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences," Bioinformatics, vol.25, no.14, pp.1722-1730, 2009.
[15] Serafim Batzoglou et al., "ARACHNE: A Whole-Genome Shotgun Assembler," *Genome Research*, vol.12, pp.177-179, 2002.
[16] William R. Pearson, David J. Lipman, "Improved tools for biological sequence comparison," *Proc Natl Acad Sci.*, vol.85, pp.2444-2448, 1988.
[17] K. Mavromatis et al., "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods," *Nature Methods*, vol.4, no.6, pp.495-500, 2007.
[18] ftp://ftp.ncbi.nih.gov/genbank/genomes/
[19] M. J. Moon et al., "A Computational Approach to Detect CNVs Using High-throughput Sequencing," *Proceeding of BIBE 2009*, pp.266-271, 2009.
[20] A. J. Iafrate, et al., "Detection of large-scale variation in the human genome, Nature Genetics, vol.36, no.9, pp.949-951, 2004.
[21] http://www.ncbi.nlm.nih.gov/projects/SNP/

부록 1 데이터셋을 구성하는 genome의 목록

Super-kingdom	Phylum	Class	Order	Family	Genus	Species
Archaea	Euryarchaeota	Methano-microbia	Methanosarcinales	Methanosarcinaceae	Methanococcoides	Methanococcoides burtonii DSM 6242
Archaea	Euryarchaeota	Methano-microbia	Methanosarcinales	Methanosarcinaceae	Methanosarcina	Methanosarcina barkeri str. Fusaro
Archaea	Euryarchaeota	Methano-microbia	Methanosarcinales	Methanosarcinaceae	Methanospirillum	Methanospirillum hungatei JF-1
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Frankiaceae	Frankia	Frankia sp. CcI3
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Frankiaceae	Frankia	Frankia sp. EAN1pec
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Kineosporiaceae	Kineococcus	Kineococcus radiotolerans SRS30216
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	Arthrobacter	Arthrobacter sp. FB24
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Nocardioidaceae	Nocardioides	Nocardioides sp. JS614
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Nocardiopsaceae	Thermobifida	Thermobifida fusca YX
Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	Bifidobacterium longum DJO10A
Bacteria	Actinobacteria	Actinobacteria	Rubrobacterales	Rubrobacteraceae	Rubrobacter	Rubrobacter xylanophilus DSM 9941
Bacteria	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Flexibacteraceae	Cytophaga	Cytophaga hutchinsonii ATCC 33406
Bacteria	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	Chlorobium	Chlorobium limicola DSM 245
Bacteria	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	Chlorobium	Chlorobium phaeobacteroides DSM 266
Bacteria	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	Chlorobium	Chlorobium phaeovibrioides DSM 265
Bacteria	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	Pelodictyon	Pelodictyon luteolum DSM 273
Bacteria	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	Pelodictyon	Pelodictyon phaeoclathratiforme BU-1
Bacteria	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	Prosthecochloris	Prosthecochloris aestuarii DSM 271
Bacteria	Chloroflexi	Chloroflexi	Chloroflexales	Chloroflexaceae	Chloroflexus	Chloroflexus aurantiacus J-10-fl
Bacteria	Cyanobacteria	-	Chroococcales	-	Synechococcus	Synechococcus elongatus PCC 7942
Bacteria	Cyanobacteria	-	Nostocales	Nostocaceae	Anabaena	Anabaena variabilis ATCC 29413
Bacteria	Cyanobacteria	-	Oscillatoriales	-	Trichodesmium	Trichodesmium erythraeum IMS101
Bacteria	Cyanobacteria	-	Prochlorophytes	Prochlorococcaceae	Prochlorococcus	Prochlorococcus marinus str. NATL2A
Bacteria	Deinococcus -Thermus	Deinococci	Deinococcales	Deinococcaceae	Deinococcus	Deinococcus geothermalis DSM 11300
Bacteria	Firmicutes	Bacilli	Bacillales	-	Exiguobacterium	Exiguobacterium sibiricum 255-15
Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	Bacillus cereus subsp. cytotoxis NVH 391-98
Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus	Enterococcus faecalis V583
Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Lactobacillus brevis ATCC 367
Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Lactobacillus casei ATCC 334
Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365
Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Lactobacillus gasseri ATCC 33323
Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293
Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Oenococcus	Oenococcus oeni PSU-1
Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Pediococcus	Pediococcus pentosaceus ATCC 25745
Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus thermophilus LMD-9
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Alkaliphilus	Alkaliphilus metalliredigens QYMF
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium	Clostridium beijerinckii NCIMB 8052

Super-kingdom	Phylum	Class	Order	Family	Genus	Species
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium	Clostridium thermocellum ATCC 27405
Bacteria	Firmicutes	Clostridia	Clostridiales	Syntrophomonadaceae	Syntrophomonas	Syntrophomonas wolfei subsp. wolfei str. Goettingen
Bacteria	Firmicutes	Clostridia	Thermoanaero-bacterales	Thermoanaero-bacterales FamilyIII.IncertaeSedis	Caldicellulosiruptor	Caldicellulosiruptor saccharolyticus DSM 8903
Bacteria	Firmicutes	Clostridia	Thermoanaero-bacterales	Thermoanaero-bacteriaceae	Moorella	Moorella thermoacetica ATCC 39073
Bacteria	Firmicutes	Clostridia	Thermoanaero-bacterales	Thermoanaero-bacteriaceae	Thermoanaerobacter	Thermoanaerobacter pseudethanolicus ATCC 33223
Bacteria	Firmicutes	Clostridia	Thermoanaero-bacterales	Thermoanaero-bacteriaceae	Thermoanaerobacter	Thermoanaerobacter sp. X514
Bacteria	Firmicutes	Clostridia	Thermoanaero-bacterales	Thermoanaero-bacteriaceae	Thermoanaerobacter	Thermoanaerobacter tengcongensis MB4
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Bradyrhizobium	Bradyrhizobium sp. BTAi1
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Nitrobacter	Nitrobacter hamburgensis X14
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Nitrobacter	Nitrobacter winogradskyi Nb-255
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodopseudomonas	Rhodopseudomonas palustris BisA53
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodopseudomonas	Rhodopseudomonas palustris BisB18
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodopseudomonas	Rhodopseudomonas palustris BisB5
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodopseudomonas	Rhodopseudomonas palustris HaA2
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	Mesorhizobium	Mesorhizobium sp. BNC1
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Jannaschia	Jannaschia sp. CCS1
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Paracoccus	Paracoccus denitrificans PD1222
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Rhodobacter	Rhodobacter sphaeroides 2.4.1
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Ruegeria	Silicibacter sp. TM1040
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Rhodospirillum	Rhodospirillum rubrum ATCC 11170
Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	Ehrlichia	Ehrlichia canis str. Jake
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	Novosphingobium aromaticivorans DSM 12444
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingopyxis	Sphingopyxis alaskensis RB2256
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	Burkholderia cenocepacia AU 1054
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	Burkholderia cenocepacia HI2424
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	Burkholderia sp. 383
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	Burkholderia vietnamiensis G4
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	Burkholderia xenovorans LB400
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Polaromonas	Polaromonas sp. JS666
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Rhodoferax	Rhodoferax ferrireducens T118
Bacteria	Proteobacteria	Betaproteobacteria	Hydrogenophilales	Hydrogenophilaceae	Thiobacillus	Thiobacillus denitrificans ATCC 25259
Bacteria	Proteobacteria	Betaproteobacteria	Methylophilales	Methylophilaceae	Methylobacillus	Methylobacillus flagellatus KT
Bacteria	Proteobacteria	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae	Nitrosomonas	Nitrosomonas eutropha C91
Bacteria	Proteobacteria	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae	Nitrosospira	Nitrosospira multiformis ATCC 25196
Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Dechloromonas	Dechloromonas aromatica RCB
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio	Desulfovibrio desulfuricans G20
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Geobacteraceae	Geobacter	Geobacter metallireducens GS-15
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Pelobacteraceae	Pelobacter	Pelobacter carbinolicus DSM 2380
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Pelobacteraceae	Pelobacter	Pelobacter propionicus DSM 2379
Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Myxococcaceae	Anaeromyxobacter	Anaeromyxobacter dehalogenans 2CP-C
Bacteria	Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophobacteraceae	Syntrophobacter	Syntrophobacter fumaroxidans MPOB
Bacteria	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	Sulfurimonas	Sulfurimonas denitrificans DSM 1251
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Marinobacter	Marinobacter aquaeolei VT8
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Saccharophagus	Saccharophagus degradans 2-40
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalteromonadaceae	Pseudoalteromonas	Pseudoalteromonas atlantica T6c
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	Shewanella amazonensis SB2B
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	Shewanella baltica OS155
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	Shewanella frigidimarina NCIMB 400
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	Shewanella loihica PV-4
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	Shewanella putrefaciens CN-32
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	Shewanella sp. ANA-3
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	Shewanella sp. MR-7 27
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	Shewanella sp. W3-18-1

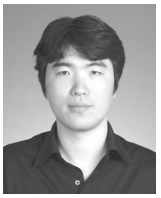
Super-kingdom	Phylum	Class	Order	Family	Genus	Species
Bacteria	Proteobacteria	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	Alkalilimnicola	Alkalilimnicola ehrlichei MLHE-1
Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Chromohalobacter	Chromohalobacter salexigens DSM 3043
Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Actinobacillus	Actinobacillus succinogenes 130Z
Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Histophilus	Haemophilus somnus 129PT
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Psychrobacter	Psychrobacter arcticus 273-4
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Psychrobacter	Psychrobacter cryohalolentis K5
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Azotobacter	Azotobacter vinelandii DJ
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas fluorescens Pf0-1
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas putida F1
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas syringae pv. syringae B728a
Bacteria	Proteobacteria	Gammaproteobacteria	Thiotrichales	Piscirickettsiaceae	Thiomicrospira	Thiomicrospira crunogena XCL-2
Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Xylella	Xylella fastidiosa 9a5c
Bacteria	Proteobacteria	unclassified	unclassified	unclassified	Magnetococcus	Magnetococcus sp. MC-1

여 윤 구

2009년 연세대학교 컴퓨터과학과 졸업(학사). 2011년 연세대학교 컴퓨터과학과 졸업(석사). 관심분야는 데이터베이스, 데이터 마이닝, 바이오인포매틱스

문 명 진

2007년 중앙대학교 컴퓨터공학과 졸업(학사). 2009년 연세대학교 컴퓨터과학과 졸업(석사). 관심분야는 데이터베이스, 데이터 마이닝, 바이오인포매틱스



김 우 철

2003년 연세대학교 컴퓨터과학과 졸업(학사). 2006년 연세대학교 컴퓨터과학과 졸업(석사). 2010년 연세대학교 컴퓨터과학과 졸업(공학박사). 2010년~현재 연세대학교 컴퓨터과학과 박사후 과정. 관심분야는 유사 검색 기법, 바이오인포메틱스, LBS

박 상 현

정보과학회논문지 : 데이타베이스
제 38 권 제 1 호 참조