

유전체 조립 방법을 접목한 올리고뉴클레오티드 빈도 기반 메타유전체 비닝 방법

여윤구[○] 문명진 김우철 박상현
연세대학교 컴퓨터과학과
{yyk, psiwind, twelvepp, sanghyun}@cs.yonsei.ac.kr

A Binning method for metagenome using Genome assembly and oligonucleotide frequencies

Yunku Yeo[○], Myungjin Moon, Woocheol Kim, Sanghyun Park
Dept. of Computer Science, Yonsei University

요 약

메타유전체는 다양한 유전체의 정보와 상호 작용에 대한 정보를 얻을 수 있는 영역이다. 메타유전체를 구성하는 생물 구성이 매우 복잡하기 때문에, 메타유전체 내에 존재하는 미생물의 종류와 비율을 알아내는 비닝 문제(Binning)가 메타유전체학에 있어서 중요한 문제 중 하나이다. 비닝 문제를 해결하기 위한 한 가지 방법으로 올리고뉴클레오티드의 출현 빈도(Oligonucleotide frequency)를 이용하는 다양한 연구가 진행되어 왔다. 그러나 올리고뉴클레오티드의 출현 빈도가 전체 유전체의 특징을 반영하기 위해서는 사용하는 유전체 조각의 길이가 일반적인 리드(read)의 길이 이상이 되어야 하는 문제점이 있다. 이와 같은 문제를 해결하기 위하여 본 논문에서는 메타유전체 특징에 적합한 유전체 조립 기법을 접목하여 유전체 조각의 크기를 증가시킨 후, 올리고뉴클레오티드 출현 빈도를 이용하는 비닝 방법을 제안하였다. 실제 메타유전체 데이터에 새로운 비닝 방법을 적용한 결과, 올리고뉴클레오티드 빈도를 이용하여 메타유전체의 유전체 조각을 더욱 효율적으로 구분할 수 있었다.

1. 서 론

메타유전체는 다양한 유전체 정보를 얻을 수 있는 정보의 보고로서 가치가 매우 높다. 메타유전체학에서는 미생물 환경에 존재하는 모든 생물의 유전체를 한 번에 추출한다. 이를 통해 연구실 환경에서 배양할 수 없는 미생물을 연구하거나, 실제 환경 내에서 미생물들 간의 상호 작용을 연구할 수 있다. 그러나 이로 인해 동시에 유전체 구성의 복잡도가 매우 높아지는 문제점이 발생한다. 거기에, 메타유전체 내에 존재하는 미생물은 같은 종이라도 하더라도 각기 다른 개체일 수 있으며, 유전체 변이(Polymorphism) 또한 심하다. 때문에 유전체 서열의 유사성에 기반을 두는 유전체 조립(Genome assembly)과 같은 방법이 큰 효과를 거두지 못했다.

이러한 어려움 때문에 메타유전체의 비닝(Binning)이 중요한 문제이자 메타유전체 연구의 시작점으로 대두되었다. 비닝 문제는 메타유전체 리드(read)를 같은 종에서 비롯된 것끼리 분류하는 문제이다. 이를 통해 메타유전체(즉, 환경) 내에 존재하는 미생물의 종류와 비율을 추정할 수 있으며, 추후 메타유전체 연구의 기반이 될 수 있다.

일반적으로 메타유전체의 비닝을 위해 가장 많이 사용하는 방법은 16S rRNA를 이용하는 방법이다. 16S rRNA

는 대부분의 미생물에 존재하는 짧은 염기 서열로서, 미생물 간의 비교를 통해 계통상의 위치를 추정할 수 있는 강력한 방법이다. 이 방법은 가장 정확하게 미생물의 존재 유무를 알아낼 수 있는 방법이지만, 생물학적 라이브러리의 구축과 실험을 별도로 수행해야 한다는 단점이 있다.

이와 같은 생물학적 실험 없이, 메타유전체 연구를 위해 구축된 염기 서열 라이브러리만을 가지고 비닝을 수행하기 위한 많은 연구가 진행되었다. 그 중 대표적인 것이 올리고뉴클레오티드의 출현 빈도(Oligonucleotide frequency)이다. 이것은 전체 유전체 내에서 n 길이의 짧은 염기 서열의 출현 빈도를 유전체의 특징 벡터로 이용하는 방식이다. 예를 들어, 테트라뉴클레오티드 빈도를 이용한다면, 전체 유전체 내에서 길이 4의 모든 염기 서열의 출현 빈도를 계산한다. 이것은 256차원(44)의 벡터가 된다. 이러한 올리고뉴클레오티드 빈도는 유전체의 특징을 나타내는 유전체 지표(Genome signature) 중 하나이다[1].

올리고뉴클레오티드 빈도가 의미를 갖기 위해서는 추출에 사용된 유전체의 크기가 충분히 커야 한다. 그래야 전체 유전체의 빈도를 유사하게 반영할 수 있기 때문이다. 그러나 메타유전체 환경에서 추출한 리드의 크기가 매우 작기 때문에, 리드에서 바로 올리고뉴클레오티드 빈도를 추출할 경우 전체 유전체의 특징을 올바르게 반영하지 못할 수 있다.

본 논문은 교육과학기술부 한국연구재단의 미래기반기술개발사업(2009-0083311)의 지원을 받아 수행되었습니다.

기존의 연구에서는 메타유전체의 리드를 직접 사용한 것이 아니라, 완전한 유전체 서열에서 큰 길이의 유전체 조각을 가상으로 추출한 뒤 사용하였다. 대부분의 연구에서 10~40Kbp로 비교적 큰 크기의 유전체 조각을 사용하였기 때문에, 유전체 전체의 특징은 잘 보존하였으나 5~700bp에 불과한 실제 메타유전체 리드에 적용하기에는 어려움이 있었다.

본 논문에서는 이러한 어려움을 해결하기 위하여, 실제 메타유전체 프로젝트의 리드 데이터에 유전체 조립 알고리즘의 일부를 적용하여 유전체 조각의 크기를 증가시켰다. 이후 크기가 증가된 유전체 조각에서 추출한 올리고뉴클레오티드 빈도를 짧은 리드에 그대로 적용했을 때와 비교 분석하였다. 그 결과 각 유전체 조각마다 올리고뉴클레오티드 빈도가 더 분명해져서, 메타유전체 내의 유전체 조각을 더 잘 구별할 수 있었다.

2. 관련 연구

Teeling 등[2]은 메타유전체로부터 40Kbp 길이의 유전체 조각을 임의로 추출하여 그것에서 테트라뉴클레오티드(tetra-nucleotide)의 출현 빈도를 조사하였다. 그 결과 서로 다른 종에서 추출한 테트라뉴클레오티드 빈도 사이에서 상대적으로 낮은 상관관계(Correlation)가 나타났으며 같은 종에서 추출한 테트라뉴클레오티드 빈도 사이에서는 높은 상관관계 값이 나타났다.

Abe 등[3]은 메타유전체로부터 1Kbp, 10Kbp 길이의 유전체 조각을 각각 추출하여 그로부터 다이(di-), 트라이(tri-), 테트라뉴클레오티드 빈도를 조사하였다. 이후 조사한 테트라뉴클레오티드 빈도를 SOM(Self Organizing Map)의 특성 벡터로 사용하여 클러스터링을 수행하였다. 그 결과 10Kbp 길이의 유전체 조각은 의미 있는 군집(cluster)을 형성하였으나, 1Kbp 길이의 유전체 조각은 군집을 형성하지 못했다.

Bohlin 등[4]은 올리고뉴클레오티드의 빈도를 분석하는 여러 가지 통계적 방법들을 비교 분석하였다. 이들은 ZOM(Zero'th Order Markov method), MCM(Markov Chain Method), ROF(Relative Oligonucleotide Frequency) 등의 방법을 이용하였으며, 2~6 길이의(di~hexa) 올리고뉴클레오티드를 사용하여 성능을 비교

하였다. 연구 결과, 사용되는 상황에 따라 다르지만 대체적으로 테트라뉴클레오티드를 사용하는 것이 성과와 메모리 등 여러 방면에서 골고루 우수한 것으로 나타났다.

본 연구팀은 이전 연구에서 5Kbp, 7Kbp, 10Kbp, 20Kbp 길이의 유전체 조각에서 각각 추출한 테트라뉴클레오티드 빈도를 이용하여, 동종 및 이종간의 상관관계 차이를 분석하였다.[5] 그 결과 7Kbp 길이까지는 동종·이종간의 상관관계 차이가 비교적 크게 나타났으며, 그 이하의 길이에서는 상관관계 값이 너무 작게 나타났다.

최근에는 Andrey. K 등[6]이 400bp 크기의 유전체 조각을 이용하기도 하였으나, 메타유전체 구성의 복잡도와 유전체의 특징에 따라 결과에 많은 영향을 받았다.

3. 제안하는 방법

본 논문에서 제안하는 방법은 유전체 조립 알고리즘의 일부를 메타유전체 리드에 적용하는 것이다. 메타유전체 리드를 그대로 사용하면 테트라뉴클레오티드 빈도가 유전체 전체의 특징을 반영할 수 없다. 따라서 유전체 조립 알고리즘을 이용하여 유사성을 갖는 리드를 모은 뒤, 충분히 크게 만들어진 유전체 조각을 이용하여 테트라뉴클레오티드 빈도 분석 방법을 적용하는 것이다.

그러나 메타유전체에 기존의 유전체 조립 알고리즘을 그대로 적용할 경우, 유전체 조립이 잘 이루어지지 않을 뿐 아니라 지나치게 많은 시간과 리소스를 필요로 한다. 특히, 메타유전체 내에 존재하는 다량의 유전체 변이로 인하여, 같은 종의 유전체 조각이 잘 조립되지 않는 문제가 발생할 수 있다.

이 때문에, 본 논문에서는 기존의 유전체 조립 알고리즘의 일부를 수정하였다. 대상이 된 유전체 조립 알고리즘은 아라크네(Arachne) 어셈블러[7]의 알고리즘이다. 아라크네 어셈블러는 k 길이의 염기 서열(k-mer)을 기반으로 하는 어셈블리 알고리즘으로서, 비교하고자 하는 염기 서열에 공통적으로 존재하는 k-mer에서부터 조립을 확장해 나간다.

본 논문에서는 아라크네 어셈블러의 오버랩 디렉션 알고리즘은 그대로 활용하되, 양 쪽 염기 서열 간의 완결된 조립 서열을 작성하는 부분을 삭제하였다. 대신 양쪽

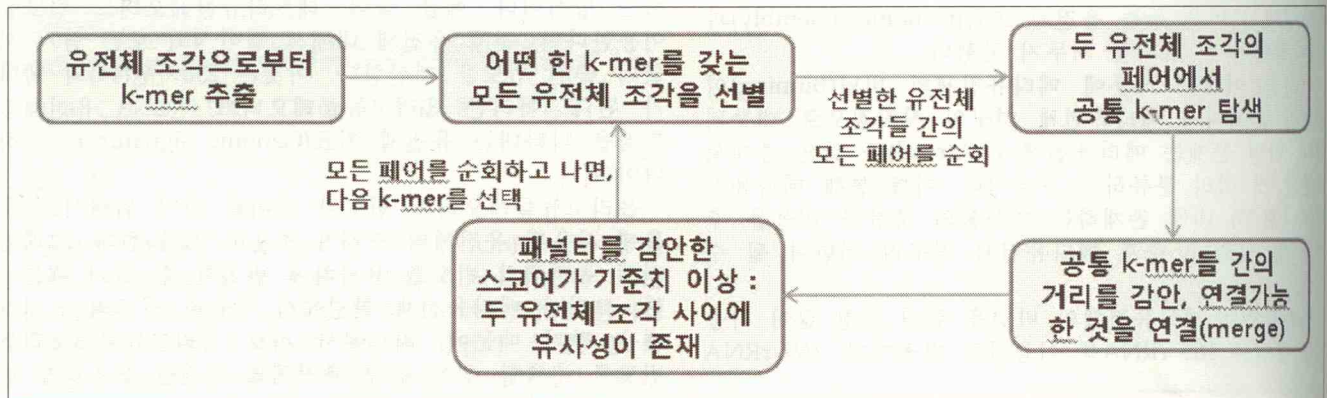


그림 1. 수정된 유전체 조립 알고리즘의 개요

