

## 유전체 특징과 유전체 조각 길이의 상관관계에 대한 올리고뉴클레오티드 빈도에 기반을 둔 비교 연구

여윤구<sup>○</sup>, 문명진, 김우철, 박상현  
연세대학교 컴퓨터과학과  
{yyk, psiwind, twelvepp, sanghyun}@cs.yonsei.ac.kr

### A Comparison of genome signature with various length genome fragments based on oligonucleotide frequency

Yunku Yeo<sup>○</sup>, Myungjin Moon, Woocheol Kim, Sanghyun Park  
Dept. of Computer Science, Yonsei University

메타유전체학(Metagenomics)은 실험실 환경에서 배양할 수 없는 수많은 미생물의 유전체에 대한 유전 정보를 얻을 수 있는 새로운 연구 방법 중 하나이다. 메타유전체학은 전통적인 유전체 연구 방식과는 다르게 특정 생물을 분리 배양한 뒤 유전체를 채취하는 것이 아니라 분리 과정 없이 환경에서 직접 유전체를 채취한다. 이를 통해 미생물 공동체(microbial community) 전체의 유전체를 한꺼번에 얻을 수 있으며, 단독으로 분리하여 연구할 수 없는 다양한 유전체 정보를 확보할 수 있게 된다.

전통적인 유전체 연구에서는 유전체 정보가 하나의 종 하나의 개체에서 생성된 정보이다. 반면 메타유전체 내에는 서로 다른 종이 섞여 있을 뿐만 아니라, 같은 종 내에서도 개체 각각의 다형성이 존재한다. 이러한 어려움으로 인하여 메타유전체 내에 존재하는 생물종의 수와 분포를 파악하는 것은 메타유전체학의 주요 목표 중 하나이다. 이를 통해 특정 환경에서 분포하는 생물종의 특징을 연구할 수 있을 뿐만 아니라, WGS(Whole Genome Shotgun sequencing)와 같은 다른 메타유전체 연구 방법에서 중요한 단서를 얻을 수 있다.

메타유전체 내에 존재하는 생물종을 추정하기 위하여, 메타유전체의 리드에서부터 바로 추론할 수 있는 유전체 특징(genome signature)을 찾기 위한 기존 연구가 수행되어 왔다[1][2]. 그 중 올리고뉴클레오티드 빈도(oligonucleotide frequency)는 종별 특이성 (species-specific characteristic)을 반영하는 유전체 특징으로 알려져 있다[3].

그러나 이러한 기존의 연구들을 메타유전체 데이터에 그대로 적용하기에는 어려움이 있다. 올리고뉴클레오티드의 빈도를 이용한 기존의 연구들은 모두 조립과 보정이 완결된 유전체 서열에서 유전체 조각을 추출하였으며, 실험에 사용한 유전체 조각의 크기도 10Kbp ~ 40Kbp로 큰 편이다. 반면 실제 유전체 리드의 길이는 평균 700bp에 불과하기 때문에 이러한 통계적인 방법을 메타유전체의 리드에 바로 적용하기는 어렵다. 그렇지만 유전체 조립 과정을 적용하여 유전체 조각의 크기를 증가시키고 메이트 페어의 정보를 이용하여 유전체 조각을 연결하면 유전체 조각의 길이를 리드의 크기 이상으로 증가시킬 수 있다. 이렇게 조립된 유전체 조각(contig)은 완결된 유전체 서열과는 달리 조립되지 못한 부분이 갭(gap)으로 남아 있으며, 상대적으로 유전체 조각의 크기도 작다.

올리고뉴클레오티드의 빈도가 유전체의 특징을 정확히 반영하기 위해서는 빈도 계산에 사용한 유전체 조각의 크기가 크면 클수록 유리하다. 반대로 너무 작은 유전체 조각을 사용하면 유전체 일부분의 특징만을 반영하게 된다. 따라서 이러한 방법을 실제 메타유전체에 적용하기 위해서는 어느 정도 길이의 유전체 조각이 유전체 특징을 유효하게 보존하는지가 먼저 검증되어야 한다. 본 논문은 올리고뉴클레오티드 빈도를 이용한 유전체의 자율적 군집화(clustering) 방법의 선행 연구로서, 유전체 조각의 크기와 유전체 특징과의 상관관계에 대해서 연구하였다.