군집화를 통한 메타유전체 종 분류 방법

박치현° 박상현
연세대학교 컴퓨터과학과
tianell@cs.yonsei.ac.kr, sanghyun@cs.yonsei.ac.kr

A Species Classification Method by Clustering in Metagenomes

Chihyun Park°, Sanghyun Park
Department of Computer Science, Yonsei University

요 약

전체 생물 중 가장 많은 비율을 차지하는 미생물을 분리 배양 시켜 연구하는 방법은 과거부터 이루어져 왔다. 하지만 다양한 유전체가 섞여 있기 때문에 실험실 환경에서 우리가 원하는 모든 유전체를 분리 배양하기란 어렵다. 최근 들어 다양한 미생물 유전체 전체를 하나의 집합으로 보고 그 자체를 분석하는 연구가 수행되고 있다. 미생물의 서식환경을 그대로 분석하는 연구는 최근 연구는 메타지놈이라고 불리고 있다. 본 논문에서는 메타지놈 환경에서 추출한 짧은 길이의 염기서열 조각을 통해 메타지놈 환경의 유전체의 종을 분석할 수 있는 방법을 제시한다. 반복적인 군집화 방법과 염기서열에서 특징을 추출하는 방법을 통해 가장 근사한 종을 수를 통계적으로 밝히며 최종적으로 하나의 메타지놈 환경에 존재하는 미생물 종의 수를 밝히는 것을 본 논문의 목표로 한다. 본 논문에서는 이 방법에 대한 전반적인 아이디어를 제시한다.

1. 서 론

미생물은 전체 생물 종에서 가장 많은 비율을 차지하고 있으며, 아직 전체 미생물들 중에서 대부분이 기능적으로, 유전적으로 밝혀진 바가 없다. 현재 실험실 환경에서 밝혀낸 미생물은 전체 미생물 종에서 약 1% 밖에 안 되는 것으로 추정되고 있다. 현대 생물학이 배양 기술과 실험 환경이 계속적으로 발전하고 있다고 하지만, 미생물의 종이 방대하기 때문에 이를 분리 배양하기란 쉽지 않다. 따라서 아직까지 미생물의 분리 배양이 쉽게 되지 않기 때문에 미생물에 대한 유전적 연구가 활발히 이루어지지 않고 있다. 미생물이 존재하는 환경은 예를 들어 하수구, 동물의 소화기관내, 지구, 심해저 해수 등으로 사실상 대부분의 자연 환경이라고 간주해도 무방하다. 예를 들어 160/ml 토양의 경우 약 6,400~38,000/g 정도의 미생물이 존재한다고 예상한다 [1]. 이는 지금까지 유전체 연구의 환경과는 완전히 다른 분포이며, 기존 대부분의 유전체 분석 방법이 통하지 않는 이유이기도 하다.

이러한 대규모 미생물들이 함께 섞여서 존재하는 유전체의 집합을 메타지놈[2] 이라고 정의 한다. 이러한 메타지놈을 연구하는 방법은 크게 두 가지로 나눌 수 있다. 실험실 환경에서 BAC-클로닝(BAC-cloning) 시스템을 통하여 미생물을 배양한 후 유전자 분석을 통해 미생물 라이브러리를 구축하는 것이 첫 번째 방법이다 [3]. 두 번째 방법은 최근 시도되고 있는 방법인데, 클로닝 시스템을 이용하지 않고 바로 염기서열을 샷건 방식으로 밝힌 후 라이브러리를 구축하는 것이다. 이른바 WGSS(Whole Genome Shotgun Sequencing)이라고 불리는 방법으로 Celera사의 Craig Venter 박사에 의해서 고안되었다. 최근 염기서열을 결정하는 이른바 시퀀싱 기술의 급속한 발전으로 실험실 환경에서의 유전체 분석을 양과 시간적으로 압도할 수 있는 기술들이 나오고 있기 때문에 최근 메타지놈 연구 연구에 도입되고 있다.

본 논문에서는 메타지놈 연구에 있어 WGSS 방식에 적용 가능한 종에 따른 유전체 분류법을 제안한다. 메타지놈에 대한 연구가 아직 초기 단계이고, 실험실 차원에서의 배양도 한계가 있기 때문에 메타지놈에 존재하는 종의 수를 확인하는 연구 또한 초기 단계이다.

현재까지 이루어진 대부분의 연구는 염기서열의 분석을 통해 메타지놈에 존재하는 미생물을 분석하기 위해서 WGS을 통해 얻어진 짧은 길이의 염기서열 조각의 특징을 추출한 후 그 특징에 따라서 메타지놈 환경에 존재하는 유전체를 추정해가는 리버스 엔지니어링 방법을 주로 사용하였다. 그렇지만 짧은 길이의 유전체 조각에서 종별 특징을 추출하는 것을 오차가 생길 수 도 있기 때문에 대부분의 연구에서는 짧은 길이의 유전체 조각을 가능한 긴 조각으로 조립한 후 그 조각들의 특징을 통해 밝혀내는 방법이 주를 이루어왔다. [4]에서는 SOM(Self Organizing Map)과 신경망(Neural Network)을 통해 2~4Bp 길이의 뉴클레오타이드에서 나타나는 유전체의 특징을