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

여윤구○, 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○, 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 뿐 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.

본 논문에서는 실제 메타유전체 프로젝트의 데이터를 이용하여5Kbp, 7Kbp, 10Kbp, 20Kbp의 유전체 조각을 생성하였다. 유전체 조각을 생성하기 위해서 실제 메타유전체의 데이터 중에서도 스캐폴드 데이터를 사용하였다. 스캐폴드 데이터는 생물학적 보정 없이 WGS의 조립 결과만을 나타내는 데이터이다. 예를 들면, 완전한 유전체 정보와는 달리 스캐폴드에는 컨티그를 조립하는 과정에서 나타나는 갭(gap)이 남아있는 상태이다. 스캐폴드에서 데이터를 추출한 것은 WGS방법에서 리드들을 조립했을 때의 중간 과정을 유사하게 재현할 수 있기 때문이다.

스캐폴드에서 추출된 유전체 조각은 256가지(4^4가지)의 테트라뉴클레오티드 단위로 출현 빈도가 계산하였다. 이후 출현 빈도를 통계적인 기댓값보다 더 나타났는지(over-represent) 덜 나타났는지(under-represent)를 나타내는 z-score로 변환하였다.[4] 이러한 과정을 통해 각 유전체 조각마다 테트라뉴클레오티드의 출현 특징을 나타내는 256차원 벡터를 생성하였다. 이후 유전체 조각의 모든 쌍에 대해서 256차원 벡터 값의 상관관계(Pearson's Correlation Coefficient)를 분석하였다.

다음으로 상관관계가 분석된 유전체 조각 쌍을 각각 같은 종에서 추출된 쌍과 다른 종에서 추출된 쌍으로 분리하였다. 이 결과를 이용하여 서로 다른 길이의 유전체 조각에서 이종 간의 유전체를 구별할 수 있는 유전체 특징이 존재하는 지 분석하였다. 그 결과, 20Kbp의 유전체 조각에서는 종내, 종간 유전체 조각 사이에서 의미 있는 결과가 나타났다. 예를 들어 같은 종 내에서 추출된 유전체 조각에서는 약 73%가 상관관계 값이 0.65를 넘는 반면, 다른 종에서 추출된 유전체 조각에서는 약 16%만이 0.65보다 큰 상관관계 값을 나타냈다. 이것은 20kbp의 길이에서는 테트라뉴클레오티드의 출현 빈도 값이 유전체의 특징을 잘 반영하고 있음을 의미한다. 10Kbp의 유전체 조각에서 역시 20Kbp보다는 다소 약하지만 의미 있는 관련성이 나타났다. 같은 종에서는 약 72%가 0.45보다 큰 상관관계 값을 나타냈지만, 다른 종에서는 약 28%만이 0.45보다 큰 상관관계 값을 나타냈다. 7Kbp의 유전체 조각에서는 같은 종에서 추출된 조각 사이에서는 62% 이상이 0.4 이상의 상관관계를 나타냈으며, 다른 종에서 추출된 조각 사이에서는 24%만이 0.4 이상의 상관관계를 나타냈다. 5Kbp의 유전체 조각에서는 같은 종에서 추출된 조각 사이에서는 54% 이상이 0.35 이상의 상관관계를 나타냈으며, 다른 종에서 추출된 조각 사이에서는 22%만이 0.35 이상의 상관관계를 나타냈다.

전체적인 실험 결과를 통하여, 유전체 조각의 길이가 길수록 유효한 유전체 정보가 보존됨을 알 수 있었다. 실제 메타유전체 프로젝트에 적용할 때에도 이런 통계적 방법을 통해 같은 종에서 비롯된 유전체 조각을 구분할 수 있을 것이다. 또한 유전체 조각의 길이가 짧아질수록 상관관계 값이 낮아지지만 모든 경우에서 같은 종 내에서 생성된 유전체 조각의 상관관계가 다른 종에서 생성된 경우보다 크게 나타났다. 다만, 유전체 조각의 길이가 7Kbp 이하인 경우에는 상관관계 값이 비교적 작게 나타났기 때문에, 통계적인 방법만을 사용하여 유전체를 구분하기보다는 다른 방법과 조합해서 사용하는 것이 적절한 것으로 보인다.

차후 이를 바탕으로 통계적인 방법만을 사용한 유전체 군집화 방법을 개발할 것이며, 이를 적용한 메타유전체 전용 어셈블러도 개발할 계획이다.

참고 문헌

[1] Takashi Abe, Shigehiko Kanaya, Makoto Kinouchi, Yuta Ichiba, Tokio Kozuki, Toshimichi Ikemura, "Informatics for Unveiling Hidden Genome Signatures", Genome Research, 13, 693-702, 2003

[2] Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, Frank Oliver Glöckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments", Environmental Microbiology, 6, 938-947, 2004

[3] David T. Pride, Richard J. Meinersmann, Trudy M. Wassenaar, Martin J. Blaser, "Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases", Genome Research, 13, 145-158, 2003

[4] Sophie Schbath, Bernard Prum, Elisabeth de Turcheim, "Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences", J Comput Biol, 2, 417-437, 1995