다염기변이 유전체에 대한 서열 정렬 툴 분석

김유선○ 김종현 여윤구 김우철 박상현
연세대학교 컴퓨터과학과
yoursun@cs.yonsei.ac.kr, angangdori@gmail.com, {yyk, twelvepp, sanghyun}@cs.yonsei.ac.kr

Analysis of sequence alignment Tools on polymorphic genomes

YooSun Kim○, Jong Hyun Kim, Yun Ku Yeo, Woo-Cheol Kim, Sanghyun Park
Department of Computer Science, Yonsei University

요 약

생명공학 기술의 발달로 지놈 프로젝트를 통해 인간・초파리 등 여러 종의 유전체 정보가 밝혀 졌다. 그러나 Post-Genome 연구에 있어서 매우 중요한 생물체인 멍게(*Ciona intestinalis*)와 성게(*Strongylocentrotus purpuratus*)의 유전체 서열은 현재 공개되어 있으나 염기서열의 연속성(continuity)에는 심각한 문제점이 존재하고 있다. 이들은 염기서열에 변이가 많은 다염기변이 유전체(polymorphic genomes)로 그 특성이 반영되지 않은 전통적인 Whole Genome Shotgun Sequencing(WGSS)방법을 사용였기 때문이다. 이와 같은 다염기변이 유전체 서열 분석은 시스템 생물학이나 비교 유전체학 등의 후발 연구에 기초가 되므로 매우 중요하다. 본 논문에서는 다염기변이 유전체에 대해 알아보고 서열 조립 알고리즘의 기본이 되는 서열 정렬 툴들 중 가장 많이 사용되는 FASTA, BLAST, BLAT에 대해 분석하여 봄으로써 다염기변이 유전체에 적합한 서열 조립 전략 수립을 위해 고려해야 하는 사항들을 논의해 본다.

1. 서 론

유전자(gene)는 유전정보의 기본단위로 A(아데닌)・C(시토신)・G(구아닌)・T(티민)으로 표시되는 4가지 염기의 배열로 이루어져 있다. 유전자는 중심원리(central dogma)에 의해 전사(trascription)와 번역(translation)을 거쳐 생명체의 형질을 발현하게 되며, 유전을 통해 그 형질을 자손에게 전달하게 하는 중요한 물질이다. 결국 유전자의 염기 배열에 의해 생명체의 형질이 결정되는 것이기 때문에 유전자의 염기서열 분석(sequnecing)은 생명체 연구에 있어 기초적이고 필수적인 작업이라 할 수 있다.

현재 유전체학(genomics)은 유전체(genome) 서열 분석을 위해 한 생명체의 유전체 정보 전체를 서열화한 다음 유전자를 예측해 나가는 접근 방법을 취하고 있는데, 이를 위해 유전체 서열 조립(sequence assmbly) 알고리즘이 사용되고 있다. 생명공학의 대용량처리(high-throughput) 기술 발달에 힘입어 유전체 조립 알고리즘에도 대용량 처리기술을 이용한 Whole Genome Shotgun Sequencing(WGSS) 방법이 개발되었는데, 이는 유전체를 무작위로 자른 작은 조각들을 염기서열 판독 기술을 이용하여 서열화하고, 이 서열에서 서로 중첩되는 부분을 연결하면서 유전체를 조립하는 방법이다[6]. WGSS 알고리즘들은 서열 정렬 툴(alignment tools)의 알고리즘에 기반한 서열 조립 전략을 세우고 있으며, 조립 결과에 대한 분석에도 정렬 툴들을 이용하고 있다. 그러므로 서열 조립 전략을 세우는데 있어 서열 정렬 툴을 이해하는 것은 필수적이다.

† 본 연구는 교육과학기술부 과학재단의 특정연구개발사업(2007-03965)의 지원을 받아 수행되었습니다.

현재까지 진행되어 온 인간이나 초파리와 같이 염기변이가 적은 유전체들의 분석은 기존의 WGSS 방법들에 의해 쉽게 조립되어 서열화된다. 그러나 우리가 살고 있는 자연계에는 수많은 미생물들을 포함하여 염기변이(sequence polymorphism)가 많은 생명체들이 훨씬 많이 존재하고 있으며, 이런 다염기변이 유전체(polymorphic genomes) 서열 분석은 post-genome연구에서 중요한 화두가 되고 있다. 대표적으로 다염기변이 유전체의 서열을 비교하므로써 비교유전체학 (comparative genomics)나 진화생물학 (evolutionary biology)을 연구하는데 기초가 되며, 미생물들이 생산하는 항생물질을 발견해 내어 신약 개발 및 질병연구 등에 기여할 수 있다. 그런데 멍게들(*Ciona intestinalis*, *Ciona savignyi*)과 성게(*Strongylocentrotus purpuratus*)와 같은 다염기변이 유전체들의 염기서열을 조립하는 데에 전통적인 WGSS 방법을 사용했을 때에 한계가 있다는 것이 밝혀졌다[8]. 또한 현재 생물정보학(bioinfomatics)의 흐름이 시스템생물학(system biology)과 같은 기능적인 방향을 지향하고 있기 때문에, 시스템생물학에서 중요한 다염기변이 생물들의 염기서열 분석의 문제점은 이러한 후발 연구의 장애요인으로 작용할 수밖에 없다. 다시 말하면 수많은 다염기변이 유전체들의 염기서열이 결정되어 가야만 이를 이용한 다른 연구들의 효율적인 진행이 가능할 것이므로 다염기변이 유전체들의 특성을 반영한 효과적인 유전체 조립 방법의 개발이 필요하다.

본 논문에서는 이와같은 다염기변이 유전체 조립 전략을 세우기 위해 기본이 되는 기존의 서열 정렬 툴들을 분석하고 이들을 다염기변이 유전체에 적용했을 때의 문제점에 대해 논의하고자 한다. 이를 위해 먼저 다염기변이 유전체와 전통적 WGSS 방법에 대해 간략히 살펴보고, 서열 정렬 툴의 알고리즘을 분석하여 이들의 특성을