참고 모델 시퀀스를 이용한 세포내 기관별 분석 방법

정재희*, 지민근‡, 정유라§, 이강만§
*홍익대학교 교양학부, ‡강릉원주대학교 컴퓨터공학과, §동국대학교 멀티미디어공학과
e-mail : jhjung@hongik.ac.kr, mingeun@cs.gwnu.ac.kr, stujung1008@hanmail.net, gangman@dongguk.edu

A method for analyzing of the organization in the cell from the whole reference genome mapping

요 약

차세대 염기 분석(NGS) 장비의 발달로 시퀀스 분석에 대한 연구는 가속화 되고 있다. 조각들로 이루어진 리드들을 어셈블하는 방법부터 이미 유전체의 정보가 알려진 데이터베이스를 이용하여 정보를 명명하는 방법까지 다양한 방법들에 대한 방법들이 주를 이루고 있다. 하지만 어셈블하는 툴마다 다른 입력 포맷을 요구하고 있어 NGS 의 결과로 다양한 방법으로 어셈블하여 비교 분석하기 쉽지 않다. 뿐만 아니라 생물 학자들이 세포내의 진화나 계통발생학적 분류를 위한 연구를 위해서 유전체 지도 완성 후 세포내의 기관별 분리 분석이 필요하나 참고 시퀀스로부터 매핑 및 기관별 분리 분석을 위해 사용목적에 따라 입력 포맷이 다른 다른 툴을 사용해야 한다. 따라서 본 논문에서는 핵, 색소체, 미토콘드리아와 같은 세포내 기관에 대한 정보를 알기 위해 최소의 정보를 입력하여 분석하고자하는 시퀀스을 입력하여, 최대한 유사하게 매칭되는 유전체를 찾아 분석하는 방법을 제안함으로써, 진핵세포내의 발생학적 연구에 도움이 되는 방법을 제안하고자 한다.

1. 서론

차세대 염기 분석(NGS)의 급속한 발달로 여러 개의 리드들을 이용하여 어셈블리하는 방법에 대한 연구가 활발 하게 진행중이다. NGS로 부터 생성된 수만 개의 리드들이 참고할 모델 유전체가 있는 경우, 유전체와의 시퀀스 유사성을 판단하는 BLAST를 이용하여 유사 시퀀스 매칭으로 컨티그를 생성 할 수 있다. 반면 모델 유전체가 없는 경우, de novo assembly[1]를 이용하여 어셈블된 컨티그를 생성한다. de novo assembly는 리드를 일정 길이로 잘라 유사성에 따라 연결하는 방법[2]이다. 이렇게 생성된 어셈블된 컨티그들을 명명 또는 분석하여 유전체에 대한 정보를 얻는 것이NGS의 큰 목적이라 할 수 있다. 이와 같은 분석으로 특정 유전체의 계통 발생에 관한 연구 및 진화 분석도 가능하다.

어셈블된 컨티그를 이용하여 유전자 정보를 알기 위하여 각 컨티그별로 웹 기반으로 된 또는 단독으로 실행되는 BLAST로 해당 컨티그의 정보를 분석 하였다. 하지만 컨티그의 사이즈가 크다거나 비교 해야 할 데이터베이스 사이즈가 클 경우, 웹 접속 시간 경과로 인한 정보수집에 어려움을 겪었고 생물학자들이 커맨드라인으로 입력을 받아 분석하기에는 어려움이 따른다. 또 다른 컨티그를 분석 하는 방법은 세포내 기관의 찾고자 하는 정보에 따라 GeneMarks [3] 또는 NCBI에서 제공하는 BLAST-genome search[4]와 같은 프로그램을 이용하는 것이다. 이 프로그램들은 특성에 따라 입력 포맷이 서로 상이하기 때문에, 컨티그 정보만으로 원하는 핵, 미토콘드리아, 색소체와 같은 정보를 분석하기가 용이하지 않다. 따라서 본 논문에서는 생물에서 유전체 분석 시 특정 유전체만 분리하여 분석하기 어렵고, 전체 유전자 시퀀스 분석 시 유전체 모두 (핵, 미토콘드리아, 식물의 경우 색소체)가 분석되기 때문에, 어셈블된 컨티그들에서, 원하는 세포 내 기관 (핵, 미토콘드리아, 색소체)을 분리[5]하여 분석을 쉽게 하고자 하는 목적이 있다.

2. 방법

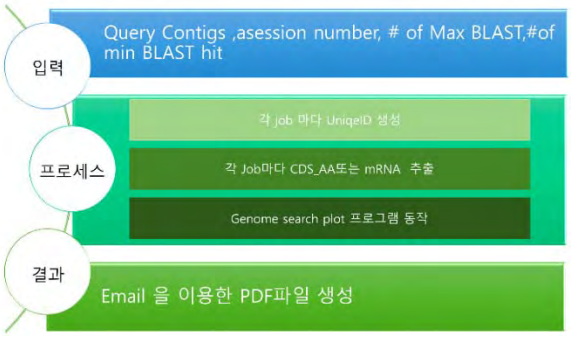

그림 1. 시스템 구성 방법

생물학자들의 편의를 위하여, 웹 기반으로 직접 컨

티그를 업로드 하여 결과를 얻는 방식으로 프로그램을 구성하였다[그림 1]. 각 분석 요청에 따른 입력이 있을 때 마다 새로운 개별 ID 를 부여하며, JOB 작업에 대한 개별 공간을 마련한다.

2.1. 입력 데이터

전제 지놈 시퀀스에서 아미노산만을 프로그램을 실행하기 위해서 어셈블된 컨티그, 두가지의 입력 시퀀스들과 두가지 옵션 값을 필요로 한다. 필요로 하는 시퀀스의 첫째는 전체 뉴클리오티드 시퀀스이고 하나는 아미노산만을 갖고 있는 시퀀스이다. 이 시퀀스는 사용자에게 직접적으로 업로드 하여 입력 받지 않고, NCBI의 accession number를 입력 받아 직접 파싱하여 시퀀스 추출물을 갖고 온다. 추출하는 방법은 NCBI의 genebank 포맷에서 CDS(Coding Sequence)라고 명명된 특징을 추출하는데 이 영역은 실제로 단백질로 발현되는 DNA 지역이다. 즉, DNA에서 Transcription, Splicing에서 Translation을 거치면서 DNA의 Intron부분은 모두 제거되고 Exon 들만 남겨진 서열을 의미한다. 필요로 하는 두개의 옵션값은 하나의 유전자 위에서 매칭되는 최대 BLAST의 결과 수를 의미하고 다른 한 개의 옵션 값은 하나의 각 컨티그에 매칭되는 최소의 부분매칭수를 의미 한다.

2.2. 프로세스

입력받은 어셈블된 컨티그를 이용하여 유전체로 BLAST를 실행한다. 각 유전체로 정렬 후 비교하는 모델 시퀀스와 쿼리 시퀀스의 CDS를 위한 아미노산 및 RNA를 위한 뉴클리오타이드를 비교한다. 비교 후 Reference sequence의 순서대로 정렬하여 각각에 gene을 명명하여 유전자 정보를 제공한다. 결과 값이 reference에 의해 정렬되어 순서가 맞지 않으면 분석하고자 하는 쿼리 시퀀스의 어셈블된 순서에 맞게 정렬 하는 것이 마지막 단계이다.

2.3. 결과

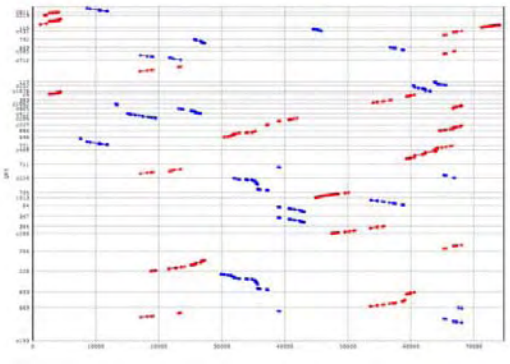

그림 2. 정렬되지 않은 컨티그들

결과는 그래프 형태의 PDF 파일로 보여주는데 x 축은 참고가 되는 모델의 시퀀스 축을 의미하고 y 축은 쿼리 컨티그가 매칭된 영역을 그래프로 표현한다. 정렬되지 않은 컨티그는 방향성을 찾기 어렵지만 [그림 2], 정렬될 경우 그림 3 과 같은 그래프를 얻을 수 있다. 표현된 그래프가 우 상향의 방향으로 되어 매칭되면 참고 시퀀스와 순행으로 매칭됨을 의미하고 좌 하향으로 매칭되어 있다면 참고 시퀀스와 역행으로 매칭됨을 의미한다.

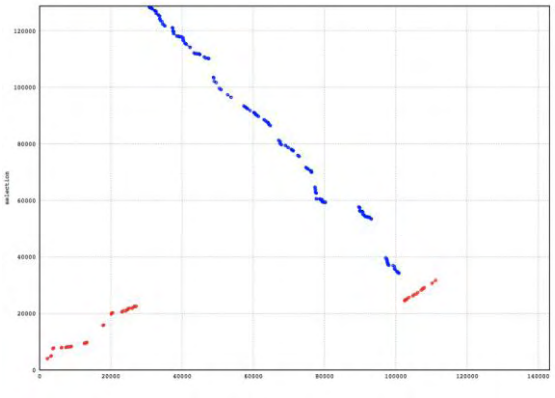

그림 3. 정렬된 컨티그

3. 결론

제안한 프로그램은 NGS 후 어셈블러로 만들어진 컨티그의 유전자 정보를 만들기 위해 만들어진 프로그램으로 참고가 될 아미노산 시퀀스과 핵산의 구성성분인 뉴클리오티드 시퀀스 어셈블된 컨티그로 유전체의 유전자 정보를 쉽게 획득 할 수 있다는 장점이 있다. 또한 여타의 다른 Plotting 프로그램들이 분석을 위해서 BLAST 와 같은 여타의 작업을 수행 하거나 입출력 값에 대한 이차적인 가공이 필요하다. 하지만, 본 프로그램은 입출력 값에 대한 재 분석이 불 필요하여 참고 가능한 시퀀스의 입력과 몇 개의 옵션값으로 질의 시퀀스에 대한 매칭되는 유전체 세포내 기관별 분석을 쉽게 할 수 있다는 장점을 갖는다.

이 논문은 연구재단의 NRF-2016R1C1B1007929 의 지원을 받아 수행된 연구임. 또한 이 논문은 이 논문은 2016 학년도 홍익대학교 학술연구진흥비에 의하여 지원되었음. 또한, 이 논문은 연구재단의 NRF-2016R1D1A1A09919318 의 지원을 받아 수행된 연구임

참고문헌

[1] Zerbino DR, Birney E (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs". Genome Res. 18 (5): 821–829.
[2] Miller JR, Koren S, Sutton G (2010). "Assembly algorithms for next-generation sequencing data". Genomics. 95 (6): 315–27.
[3] Besemer J, Lomsadze A, Borodovsky M (2001). “GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.” Nucleic Acids Research 29, pp 2607-2618
[4] https://www.ncbi.nlm.nih.gov/orffinder/
[5] Kim JI, Yoon HS, Yi G, Kim HS, Yih W, et al. (2015) “The Plastid Genome of the Cryptomonad Teleaulax amphioxeia”. PLOS ONE 10(6)