사용자 편의성 증진을 위한 유전체 분석 파이프라인 설계 및 구현

Implementation of Genome analysis pipeline for user friendly

정 민 석, 김 동 욱*, 최 한 석**
국립농업과학원 농업생명자원부, 주식회사 올비티*,
목포대학교 컴퓨터공학**

Jung Minseok, Kim Dong-Wook*, Han Suk Choi**
National Insitute of Agricultural Science, AllBT co. Ltd*,
Department of Computer Engineering Mokpo National University**

요약

본 연구는 유전체 분석 연구 중에 중요한 어셈블리 및 분석 시스템을 활용하는 데 있어 많은 연구자들이 컴퓨터 지식의 부족으로 인하여 신속한 분석을 수행하지 못하여 연구에 많은 어려움을 겪고 있어 이를 극복하기 위하여 일반 연구자도 쉽게 유전체를 분석할 수 있도록 다양한 분석방법을 자동으로 제공해주는 사용자 관점의 분석 파이프라인을 설계하고 구현하였다.

Ⅰ. 서론

게놈이란 한 생명체가 지닌 유전자의 전체를 말하며 생명현상의 유지 및 모든 내 외부 형질 발현에 필요한 유전자 정보의 데이터베이스이다. 현재 기술의 발달로 인하여 3조원이 소요되었던 게놈확보 비용이 500만원이하로 저렴해졌다. 그러나 시퀀싱 비용 하락이 많은 데이터를 양산하게 되고 이를 처리하기 위한 컴퓨터 인프라 확보 및 분석 기술의 복잡성이 대두 되어 유전체 분석의 병목현상을 일으키고 있다. 본 논문에서 제안하는 시스템은 유전체 연구자들이 쉽게 게놈 데이터를 분석, 검토, 판별할 수 있도록 사용자 편의에 중점을 둔 시스템 개발을 수행하였으며, 직관적인 그래픽 사용자 인터페이스에 대한 요구를 수용하여 사용자 관점의 인터페이스를 다양하게 구현하였다. 뿐만 아니라 생물학적 의미론을 분석하는 단계에서 이들 간의 유기적인 데이터 교류를 위한 포맷 변화를 위해 스크립트언어를 이용하여 프로그램 결과의 파싱 및 입력자료를 자동으로 생성하고 수행되도록 하였으며 이러한 과정을 대쉬보드를 통해 확인할 수 있도록 시스템을 설계하고 구현 하였다.

Ⅱ. 관련연구

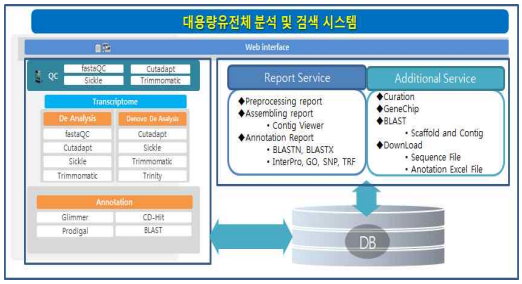
유전체 데이터를 분석하여 의미 있는 정보를 찾는 과정은 아무나 할 수 없는 어려운 작업이다. 널리 쓰이고 있는 상용프로그램들은 단일 스텝별로 제공되고 있으나 각각의 입력값과 결과값의 포맷이 상이하여 일련의 분석과정을 수행하는데에는 결과값의 파싱 및 다음 분석을 위한 파일생성을 사용자가 직접 진행하여야 하는 어려움이 있다[1]. 또한, 로컬에서 사용할 수 있는 공개된 분석 프로그램은 대부분 CUI 환경으로 설치 및 세팅이 용이하지 못한 실정이다[2].

Ⅲ. 시스템 설계 및 구현

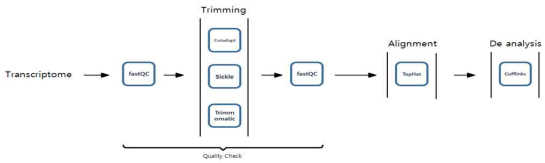
시퀀싱 기술의 발달로 인하여 샘플만 있으면 관련 업체에 의뢰하여 누구나 쉽게 시퀀싱 데이터를 얻을 수 있다.그러나 데이터의 크기가 수십기가에서 수십테라에 이르고 분석시간이 적게는 하루에서 몇 달 이상 걸리는 분석과정은 일반 생물학연구자들이 쉽게 처리하지 못하는 문제가 있다.

본 논문에서는 이를 해결하기위해 생물학 연구자들이 쉽게 접근할 수 있고 몇 번의 클릭만으로 제시된 분석과정을 수행할 수 있는 사용자 중심의 유전체 분석 파이프라인을 설계하고 구현하였다.

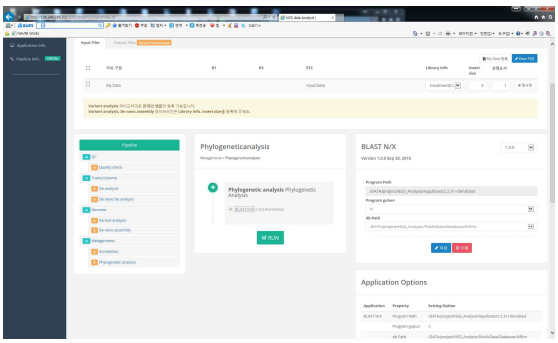
구현된 시스템은 유전체 분석을 프로세스와 프로그램 및 사용자관리를 수행하는 Pipeline Management와 Transcriptome 분석, Genome 분석, 메타게놈 분석을 수행하는 4가지 모듈로 설계되었다. 각각의 모듈별 분석방법은 국내외 논문을 참조하여 유전체 연구방법론에 맞는 분석전략을 수립하였고 이에 따라 분석이 수행되도록 20여가지의 공개 분석 프로그램을 서버에 설치 하였으며 파이프라인을 연결하기 위하여 개개의 프로그램에서 나오는 결과 값을 python을 이용하여 다음 step의 입력값이 될 수 있도록 스크립트를 이용하여 데이터를 파싱하고 변경시켜 분석이 수행되도록 시스템을 구축하였다.



▶▶ 그림 1. 대용량유전체 분석시스템 구조



▶▶ 그림 2. Transcriptome De Analysis method



▶▶ 그림 3. 시스템 사용자 인터페이스

시퀀싱된 유전체 서열을 이용하여 분석을 수행하기 위하여 본 논문에서 제시하는 시스템은 다음과 같은 step으로 시스템을 사용할 수 있다.

○로그인 후 웹 페이지 좌측에 Projects에 Data를 선택
○기존에 만들어진 폴더를 선택 또는 새로운폴더 생성
○ "자료등록" 버튼을 클릭하여 데이터 특성에 맞게 파일을 첨부, metagenome은 fasta 파일 input을 기본으로 하며, paired-end 개념과 다르기 때문에 etc로 등록
○ 웹 페이지 좌측에 Projects 선택
○ "신규" 버튼을 클릭하여 프로젝트 생성 하고 저장
○ Projects 목록 중에 새롭게 생성한 프로젝트를 선택
○ Input Files 창 오른쪽 우측에 "My Data 등록"을 클릭하여 앞서 업로드한 데이터를 선택
○ "Data 저장" 버튼 클릭
○ 화면 하단에 Pipeline에서 Metagenome 카테고리의 "Phylogenetic analysis"를 클릭
○ Annotation에 사용할 tool인 BLAST를 선택하고, 사용할 database 종류를 선택하고 저장
○ "RUN" 실행
○ Pipeline이 실행되면 Projects 상세 페이지에서 상단에 해당 프로젝트의 상태가 "RUNNING"으로 표시되고 하단에 "Project Log"를 클릭하면 pipeline에 포함된 프로그램들의 실행 log 확인 가능, 프로젝트의 상태가 "COMPELTED"로 바뀌면 "Output Files"에서 결과 파일들 다운로드 가능하여 유전체 분석을 완료할 수 있다.

Ⅳ. 결론

본 논문은 변화하고 있는 유전체 분석 시장에서 국내 최초로 유전체 Data를 효율적으로 저장하고 분석할 수 있는 시스템을 설계하고 구현하였으며 구현된 시스템을 이용하게 된다면 급격히 발전하는 유전체 분석시장을 선도 할 수 있으리라 판단된다. 향후에는 본 시스템에 딥러닝 기술을 적용한다면 지능형 유전체 분석을 수행할 수 있어 최적의 유전체 조합을 찾는데 유용할 것이며 이에 대한 연구가 필요할 것으로 생각된다.

본 성과물은(논문, 산업재산권, 품종보호권 등)은 농촌진흥청 포스트게놈다부처유전체연구사업 (PJ013693012018)의 지원에 의해 이루어진 것임

■ 참 고 문 헌 ■

[1] Aurrecoechea C, Brestelli J, Brunk BP, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, et al: EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res 2010, 38:D415-419.

[2] Ewing B, Hillier L, Wendl MC, Green P: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998, 8:175-185.