

# 21

# 긴 게놈 서열을 위한 대화형 시각화 시스템

소속 정보컴퓨터공학부

분과 B

팀명 Byte Me

참여학생 카즈타에바 굴나즈, 예르자노프 지네덴

지도교수 조환규

## 소개

### Introduction

게놈 시퀀싱 기술의 발달로 수 많은 생물체에 대해 대량의 서열 정보를 쉽게 얻을 수 있게 되었다. 그러나 시퀀싱 결과로 얻은 DNA 조각들은 길이가 짧고 불완전 하기 때문에, 다른 생물체 종의 서열 정보를 참고하거나, 또는 주어진 서열 조각의 유사한 영역 정보를 바탕으로 이어 붙여서 충분히 길이가 긴 서열이 되어야 비로소 가치가 생긴다. 이 작업은 사용자가 수 많은 바이오 소프트웨어를 설치하고, 기능을 파악해서 데이터를 분석해야 하므로 까다로운 일이다. 본 프로젝트는 사용자가 이 모든 과정을 손쉽게 분석할 수 있도록 interactive web application을 개발하는데 초점을 맞췄다.

### Scaffolding

시퀀싱 결과로 얻은 DNA 조각은 assembly 작업을 통해 contig단위의 서열 되는데 여전히 길이가 짧고 정보가 불완전하다. Scaffolding은 이러한 contig들을 결합하여 scaffold 단위의 완전한 연속된 서열로 결합하는 과정이다. 이처럼 서열 정보를 연결하는 작업은 각 과정별로 적절한 프로그램을 통한 분석을 필요로 하기 때문에, 번거롭고 시간이 많이 걸린다.

### Project Goal

본 프로젝트는 사용자가 별도의 프로그램 설치 없이도 게놈 서열 데이터를 분석하고, 그 결과를 시각화된 형태로 확인할 수 있는 interactive web platform을 개발하는 것을 목표로 한다.

## 시스템 구성

#### Web-page

사용자는 서열 데이터(fasta) 또는 서열에 대해 일부 분석된 데이터(paf, mum)를 선택해서 업로드할 수 있다



#### Linux Server

업로드 된 파일 정보를 workspace에 저장하고 1차적으로 통계 분석해서 서열 정보가 유효한지, 중복 파일인지 등을 확인한다



#### MySQL Database

작업 파일들의 조회 및 관리를 위해 분석이 완료된 파일의 정보를 database에 등록한다



#### DotPlotly

분석된 결과를 바탕으로 plot 형태의 시각화 결과를 생성한다

#### Display

사용자가 접속한 웹페이지에 시각화 결과를 출력하고 옵션을 적용할 수 있도록 한다

#### Alignment Algorithm

minimap2, mummer, chromeister 프로그램을 이용해 유사 구간을 분석한다

#### Export

서열 데이터 분석 결과 및 시각화 생성파일 등 workspace의 작업 결과물을 사용자가 다운로드할 수 있다

우리 시스템은 웹 사이트에 업로드된 서열 (분석) 데이터를 바탕으로 서열 alignment 및 시각화 알고리즘을 적용해서 서열 분석 데이터 및 시각화 데이터를 생성한다. 웹 사이트에서 시각화 결과 및 옵션을 적용한 결과를 확인할 수 있으며, 결과 파일을 다운로드 받아서 저장할 수 있다.

## 결과 및 응용

### ACHIEVED GOALS



#### 결과

우리는 사용자의 연구 과정을 간소화하고, 즉각적인 분석 결과를 시각화 형태로 제공하고자 노력했으며, 그 결과 다양한 종류의 게놈 데이터 파일에 대한 분석 및 시각화를 성공적으로 수행하는 웹 플랫폼을 만들 수 있었다. 또한 분석 결과를 데이터베이스에 등록해 정보를 효율적으로 관리할 수 있도록 하고, 사용자가 언제든지 웹페이지에서 결과를 다시 조회할 수 있도록 했다. 이 플랫폼을 통해 사용자는 번거로운 과정 없이 파일을 업로드하는 것만으로 분석 및 시각화를 얻을 수 있어 효율적으로 연구를 진행할 수 있다.

#### 향후 개선방향



향후에는 아래와 같이 시각화, 분석, 데이터베이스 기능을 보완하여 더욱 심화된 분석 결과를 제공하고, 시스템의 완성도를 높이려 한다.

- 시각화된 plot의 layout 옵션 추가
- 유사 영역 결과에 대해 order 알고리즘 적용 및 평가
- 사용자별 세션 관리

