11조 설계과제 중간 보고서 요약

2018112051 김도엽 | 2018112053 황종익 | 2018112069 김준섭 | 2018112061 남동호

- 전제

Given M number of short reads of length L, reconstruct the original sequence of length N that those shorts reads come from.
***N(Original sequence의 길이) : 30억 개**
***M(short reads의 개수) : 2천만 개**
***L(short reads의 길이) : 32**
***D(하나의 read에서 발생하는 mismatch의 개수) : 4개 까지**
***기본 형태는 paired – end reads(간격 : 40)로 주어지며 이를 활용할 수 없는 알고리즘에서는 single – end read로 받아들인다.**
데이터 생성은, 아호 코라식을 사용하여 rand 알고리즘을 통해 생성되는 substring의 반복 주기를 최대한 늘려 주기성을 최대한 제거해 주었다.

- 맡은 알고리즘 및 methods

1. 김도엽 : de novo – Velvet / KMP 이용
Velvet은 짧은 단어를 기반으로 하는 그래프를 이용해 많은 short read와 read pairs들을 이용하는 새로운 접근 방식이다.
2. 황종익 : Multiple Bloom Filter를 이용한 pattern string matching
원소가 집합에 속하는지 여부를 검사하는 확률적 자료 구조를 이용해 dna sequencing data의 패턴 스트링 매칭을 진행한다.
3. 김준섭 : SHARCGS / KMP 이용
SHARCGS은 높은 정확도로 sequence를 구성해내고 속도와 정확성적인 측면에서 기존 알고리즘을 능가한다고 알려져 있다.
4. 남동호 : reference / BWT 이용
Reference DNA와 BWT 스트링 매칭을 이용하여 얼마나 효율적인 알고리즘 성능이 나올지를 알아보려고 함.
이 위에 설명한 방법들을 이용하여 설계 프로젝트를 진행해 보려고 한다.

- 결론 도출 방식

1. 실행 시간 : 각자 맡은 알고리즘을 실행 및 변경, 보완을 하고서 실행 시간을 측정한다.
2. 각자 맡은 알고리즘의 특징을 간단히 정리 후 설명
3. 알고리즘의 시간 복잡도를 계산한 뒤, 실행 시간과 비교를 해보며그에 얼마나 비례하게 작동했는지 확인해본다. 만약, 일치하는 정도가 낮다면, 그 원인 또한 분석해본다.
4. 공간복잡도 또한 구해본다. 정확도, 정확도 / 실행 시간 등을 구해 본다.
5. 동일한 실행 환경 (windows, i7-8700k, 32gb, 1080ti)