9조 활동 보고서

6월 04일 활동 보고서

과 목	알고리즘 3분반
교 수	주종화 교수님
제 출 일	6월 4일
조 이 름	9조
조 장	2022110151 이주연
조 원	2023111033 김태은
	2021112504 박지우
	2022113556 정태호

목차

1. 향후 일정 정리

6월 3일 (화)	개별 구상한 BWT 기반 변형 알고리즘 공유
6월 6일 (금)	코드 구현 완료
6월 7일 (토)	PPT 제작 시작 – 각 구현 알고리즘 구조 및 흐름 설명 (의사코드 or 순서도) – 기존 알고리즘에서 변형된 부분 강조 설명 – 세 가지 척도(1. 최악 시간복잡도 2. 제자리성(공간복잡도) 3. 안전성)로 성능 분석 결과 삽입 – 장/단점 비교 (Trivial vs 각 조원별 작성한 알고리즘)
6월 8일 (일)	PPT 및 코드/파일 제출
6월 11일(수)	발표

2. 교수님 피드백 요약

2-1. 문제 정의 수정

단순 친자 판별 → 아이의 DNA sequence 완전 복원 후 부모와 비교하는 방향으로

아이를 기준점으로 일치율 계산 → 부모 기준으로 일치율 계산

N=3,000,000, M=10,000 고정 → 직접 실험을 통해 최적값 도출

2-2. 알고리즘 설계

기존: BWT 기반 변형
변경: Trivial 알고리즘을 벤치마킹 알고리즘으로 (BWT보다 구현 용이)

2-3. 실험 설계

단일 아이 대상 실험에서 **다수의 아이**로 범위 확장

예시: 백만 명의 아이로 친자를 검사함 → 그중 78% 이상 일치시 친자

성능 분석 그래프 포함시키기

X축: M (short read 개수)

Y축: N (sequence 길이) 와 같은 식

2-4. 발표 및 제출 유의사항

내용을 명확하게 전달하는 것이 최우선

자신이 한 부분이 명확히 구분될 것.

의사코드 금지 → 반드시 실제 프로그램 구현 후 실행 결과 도출

문제 정의부터 해결 과정까지의 논리적 흐름을 명확히 제시할 것

3. 피드백을 바탕으로 문제 정의 수정

Problem: Given 10,000 number of short reads of length 32, reconstruct the original sequence of length 3,000,000 that those shorts reads come from.

항목	설정값	설명
Reference 길이 N (부모 sequence 길이)	30만 → 300만 → 3000만 등 단계적으로 조정하여 알고리즘별 속도·정확도 변화 관찰 필요	메모리 허용 범위 내에서 알고리즘별 성능 차이가 뚜렷하게 드러나는 값으로 자체적으로 선택
Short Read 수 M (한 아이로부터 추출된 short read 개수)	1,000 → 10,000 → 100,000… 같은 식으로	아이로부터 추출한 short read의 개수, N과 동일하게 자체적으로 최적값 도출 필요, 각 단계에서 실행 시간 및 메모리 사용량 기록.
Short Read 길이 L	32	Short read 하나당 길이
허용 mismatch 수 D	3개 이하	한 개의 child 시퀀스(재구성 결과)와 parent 시퀀스 간 mismatch 개수를 통해 % 계산
레퍼런스	실제 유전자의 일부	3백만 길이 dna를 자체적으로 생성해 사용

문제 상황 정의

한 아이의 short read 데이터를 이용하여, 1. 부모 시퀀스로부터 아이의 시퀀스를 재구성(reconstruct)하고, 2. 재구성된 자식 시퀀스가 부모 시퀀스와 얼만큼 일치하는지를 계산, 친자 여부를 판

별.

부모 DNA Sequence

AGCTTAGTGATCTTTAGCCC
TAGTCACCTAGTCTCCATCA
TGCTACGGCTAAGCATATTA
TCATGATACAGGTATTC.....

.....CATG

길이 : N = 3,000,000

아이 한 명의 DNA sequence로부터

랜덤 추출된 M=10,000개의 short reads

8개 | 8개 | 8개 | 8게 **1st short read**

8개 | 8개 | 8개 | 8게 **2nd short read**

8개 | 8개 | 8개 | 8게 **3rd short read**

8개 | 8개 | 8개 | 8게 **M=10,000th short read**

허용 mismatch 수 D <= 3개

(short read를 4파트로 나눴을 때 4파트 중 한 파트는 완전히 일치하는 파트)