알고리즘

유전자 복원 프로젝트 결과 보고서

주종화 교수님
2018.06.12
2016112177 서보미

ⅰ) 프로젝트 목표:

길이 L의 short read가 M개 주어졌을 때, short read를 사용하여 길이 N 의 유전자를 복원

ii) 설계 개요

MyGenome: 길이 N 의 원본 유전자
ReferenceGenome: MyGenome과 5%정도 불일치가 일어나는 길이 N의 비교 유전자
Short Reads: MyGenome의 임의의 위치에서 가져온 길이 L의 문자열
Mismatch: Reference Genome과 MyGenome 이 완벽히 같지 않기 때문에 reference Genome에서 short reads 검색 시 어느 정도의 불일치는 허용해 준다.

iii) 사용한 알고리즘

-벤치마크 알고리즘:

○ Trivial Search: 직선적 방법으로 검색하는 알고리즘.

-사용 알고리즘:

○ BWT string matching: 기존의 BWT알고리즘은 주어진 문자열의 길이가 N일 때 NxN 행렬을 만들어 정렬 후 검색을 시작하므로 메모리 낭비가 많았다. 나는 정렬시 일부분의 문자열만 있어도 된다고 생각하여 Nx30 행렬을 생성후 정렬하였다..

-시간 복잡도: O(nm)

○ Indexing 알고리즘: 기존의 indexing 알고리즘은 Phone Book 방법을 사용할 때 이용 되었는데, Phone Book 또한 메모리 차지가 심하고, 필요에 비해 너무 많은 종류의 sequence가 있다고 생각하였다. 따라서 가능한 모든 조합의 문자열을 sequence로 사용하는 것이 아닌, short reads를 sequence로 사용 하였다.

-시간 복잡도: O(log n)

iv) 결과 분석

N: 10000, mismatch=3

short read	Trivial	BWT	Hashing
long = 20 num = 1000	132.348 87.21%	87.651 86.93%	0.205 86.29%
long = 30 num = 1000	133.234 94.47%	92.47 93.56%	0.272 91.92%
long = 30 num = 1500	193.576 98.57%	130.581 98.42%	0.184 93.22

N: 100000, mismatch=3

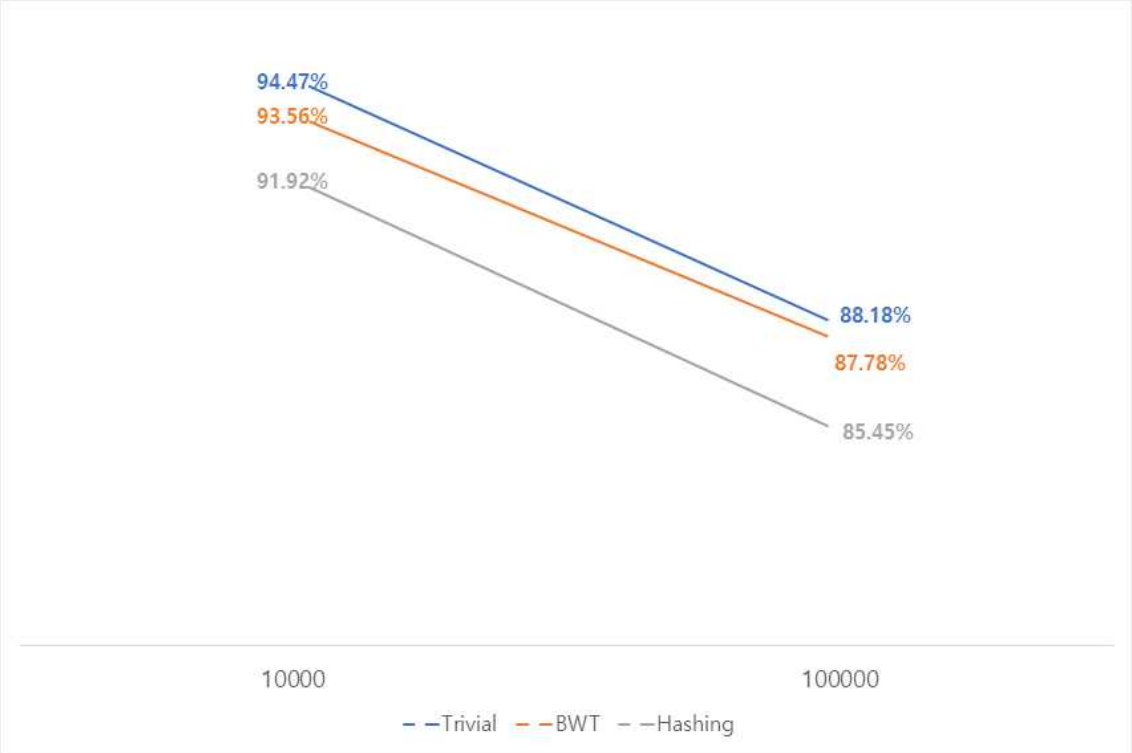
short read	Trivial	BWT	Hashing
long = 30 num = 10000	14277.439 88.182%	9821.89 87.784%	1.991 85.452%

N: 1000000, mismatch=3

	Hashing
short read long = 35 short read num = 100000	21.366 97.4212%
short read long = 40 short read num = 100000	25.708 98.1856%

시간 비교

정확도 비교



ⅴ) 개발 환경

CPU: Intel(R) Core(TM) i5-7500 CPU @ 3.40GHz
RAM: 8.00GB
OS: 64bit Window10