시험문제풀이 및 설계과제

Lecture 11

시험문제 풀이

- 마지막 문제는 알고리즘 과목 공통 문제로 알고리즘 01분반이 아직 시험을 치루지 않은 관계로 정답을 지금 공개하지 않겠습니다.

practice 7 의 문제 2

스트링 매칭은 실제로 DNA 염기서열 분석의 다양한 문제에 자주 쓰입니다. DNA 염기서열은 A/C/G/T 네개의 문자로 이루어진 서열입니다. 배운 스트링매칭 알고리즘을 실제 문제에 응용하여 봅시다.

- A/C/G/T로 이루어진 n=1000,000개의 문자를 random하게 생성하여 input.txt 파일에 저장하는 코드와 input.txt를 제출하시오.

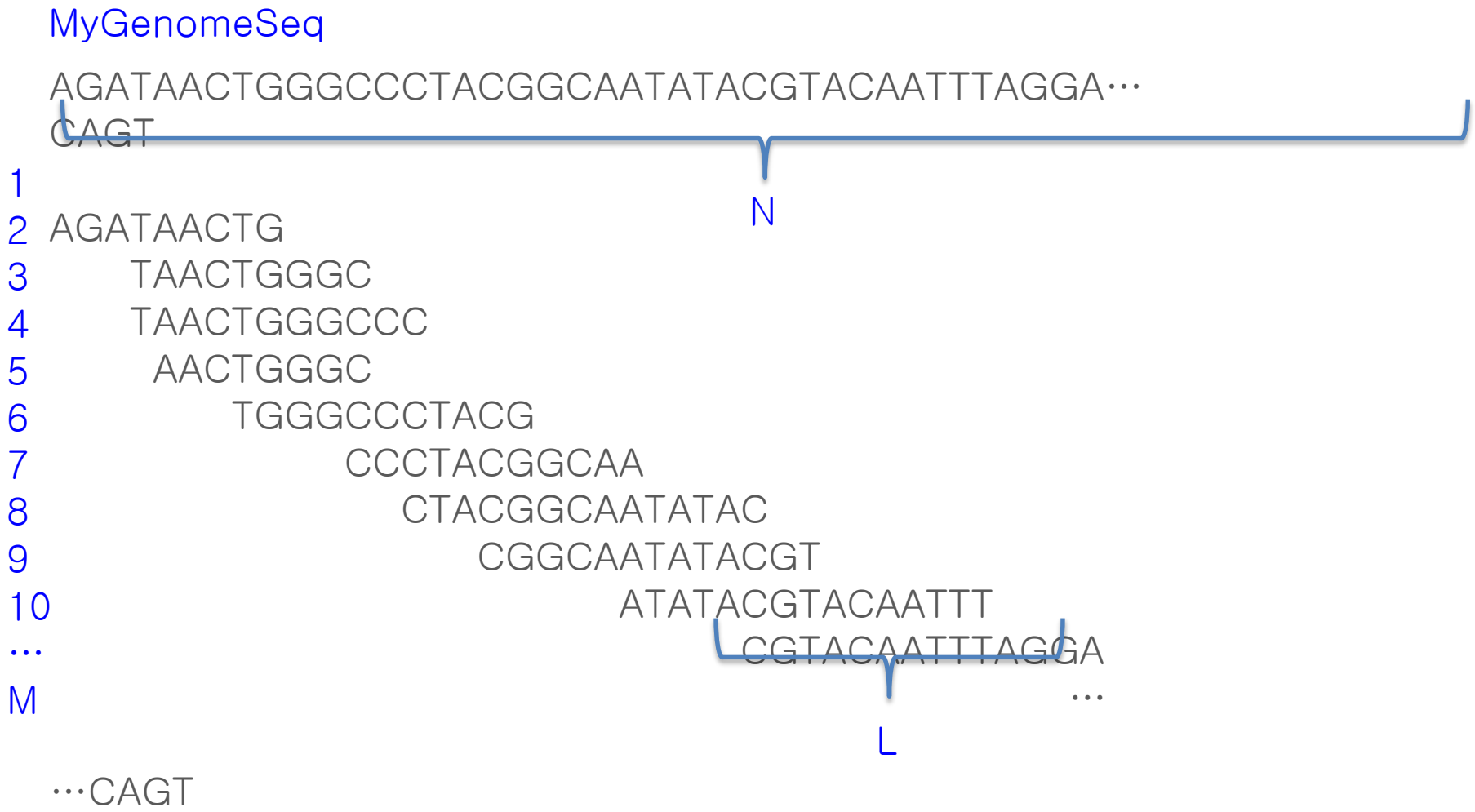
이는 이후 과제에 이용할 것입니다.

input.txt 의 예시

문제

- practice 7 의 문제 2번에서 A/C/G/T로 이루어진 길이 N= 1,000,000 짜리 sequence 를 random 하게 생성하는 것을 해보았다. 이 sequence 를 MyGenomeSeq 이라고 하자.
- 문제 1. MyGenomeSeq를 임의의 위치에서 길이 L=50 짜리로 잘라서 short read 를 만들어 보자. 이와 같은 short read 를 M=100,000 개를 만들어 reads.txt 파일에 저장하는 프로그램을 작성하여 코드와 reads.txt를 제출하시오.

예시


MyGenomeSeq
AGATAACTGGGCCCTACGGCAATATACGTACAATTTAGGA…
CAGT
N
1
2 AGATAACTG
3 TAACTGGGC
4 TAACTGGGCCC
5 AACTGGGC
6 TGGGCCCTACG
7 CCCTACGGCAA
8 CTACGGCAATATAC
9 CGGCAATATACGT
10 ATATACGTACAATTT
… CGTACAATTTAGGA
M …
L
…CAGT

문제

- Coverage 란 한 site (MyGenome의 각 문자)를 커버하는 short read의 갯수를 의미한다. 아래의 예에서 site 4의 coverage 는 3이고 site 8의 coverage 는 5이며 site 13의 coverage는 3이다.

문제

- 문제 2. 문제 1 에서 MyGenomeSeq 의 평균 coverage(N개 site coverage의 평균)는 얼마인가?

설계 문제 추가 설명

- 쉽게 설명하여 reads.txt가 주어졌을 때 input.txt의 MyGenomeSeq를 만드는 것이 이번 설계 과제의 문제입니다.

- 설계과제의 아주 간단히 예시를 준다면

입력은 acact cgaca gacac acact acgac cgaca gacac acact
원하는 출력은 acgacact

- 얼마나 acgacact를 정확하게 빠른시간내에 만드는가가 알고리즘의 성능 평가 척도가 됩니다.
- 여기서 N=8 L=5 M=8 인데 실제론 훨씬 더 긴 sequence 를 가지고 설계과제를 만들면 됩니다.