스트링 매칭

Practice 6

시작하기 전에

- 다음 실습전날 저녁 11:59시까지 eclass에 제출하시오.
- 결과물은 가능한 워드로 제출하고 코드 등 이외 제출할 것이 있는 경우 모두 압축해서 하나의 폴더로 제출할 것.
- 제출 파일 또는 폴더의 이름은 학번_이름.xxx로 할 것.

시작하기 전에

- 프로그램 작성하는 문제는 프로그램도 함께 제출하여야 함.
- 코드를 짤 때 코드의 맨 위에 자신의 학번 이름을 주석으로 적고 각 단계에 대하여 자세한 주석 달기.
- 코드를 돌려서 결과물을 제출하라는 문제는 결과물을 제출할때 화면 캡쳐를 사용할 것. 이는 자신의 코드를 돌렸을때 나온 결과임을 보이기 위함으로 사용하는 언어나 에디터 등등에 따라 다를 수 있으므로 방법은 알아서 제출할 것. 어떤 방식이든 자신의 코드를 돌려서 나온 결과라는 것만 보여주면 됨. 예를들어, Visual Studio를 사용하면 이때 자신의 코드의 윗부분(학번 이름과 앞에 코드 5줄정도 포함)이 실행 결과와 같이 캡쳐되도록하시오.

문제 1. 스트링 매칭 알고리즘 구현

- Text String과 Pattern String을 입력받아 1차원 배열에 저장하고, Text String 에서 Pattern String 을 모두 찾는 프로그램을 Brute-Force, KMP, Rabin-Karp 알고리즘을 이용하여 작성하시오.
  - 각 알고리즘의 코드를 제출하시오.
  - 아래의 Text String 과 Pattern String을 입력한 후 각 알고리즘의 결과와 수행시간(단위는 적절하게 알아서)을 출력하시오.

Text String: A STRING SEARCHING EXAMPLE CONSISTING OF A GIVEN PATTERN STRING
Pattern String: STRING

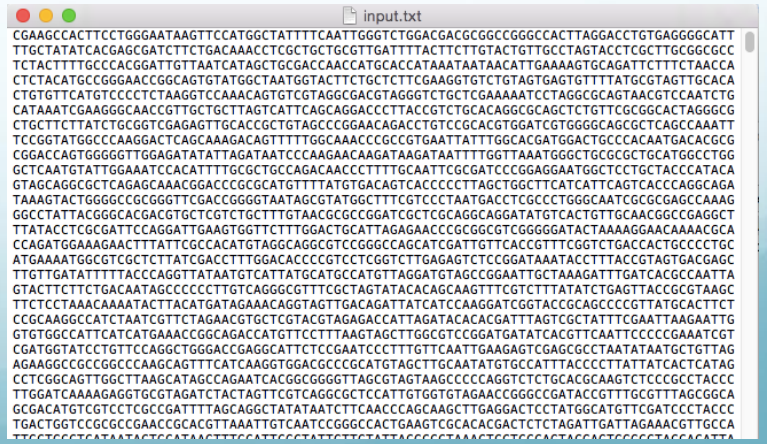
문제 2. 스트링 매칭 알고리즘 구현

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=100,000개의 문자를 random하게 생성하여 input.txt 파일에 저장하는 프로그램을 작성하여 코드와 input.txt를 제출하시오.

이는 이후 과제에 이용할 것입니다.

input.txt 의 예시 →



문제 2. 스트링 매칭 알고리즘 구현

- input.txt에서 랜덤한 A/C/G/T로 이루어진 길이 m 패턴을 모두 찾아 각 인덱스(인덱스는 0부터 시작)를 output.txt 파일에 출력으로 저장하는 프로그램을 직선적 방법(A) 과 그 이외의 알고리즘을 사용하여 구현(B) 하시오.
  - 수업시간에 배운 알고리즘을 응용하거나 그 외에 자신이 원하는 어떤 방식으로 구현해도 됩니다. 다만 적어도 특정 상황에서 직선적 방법보다는 어떤 면(시간, 메모리, 정확도 등)에서의 성능이 좋아야 합니다. 예를 들어 내 알고리즘은 메모리를 많이 사용하는 대신 n이 얼마 이상이고 m이 얼마 이상 일때 (A)방법 보다 수행시간이 적게 걸린다. 구현한 알고리즘과 장단점을 자세히 분석하여 설명하고 코드에 자세하게 주석을 달아 제출하시오.

문제 2. 스트링 매칭 알고리즘 구현

- (A), (B) 프로그램 두개의 코드와 n=100000, pattern = ACCGTAT일 때의 input.txt과 output.txt을 제출하시오.
- n을 1000부터 10배씩 증가해가면서, 각각에 대해 m을 5에서 30까지 적당히 (예를 들어 m=5,10,15,20,40) 늘려가면서 수행시간(시간단위는 비교가 보여질 수 있는 정도로 적절하게 각자 정할 것)을 측정하고 이를 직선적 방법으로 구현한 것과 비교하여 그래프로 그리시오. 이때 수행시간이 너무 길어지면(예를 들어 30분이상) n을 그만 증가시켜도 됩니다. 그래프를 그리는 방식은 각자 고민해 보고 알아 보기 편하도록 그리시오.