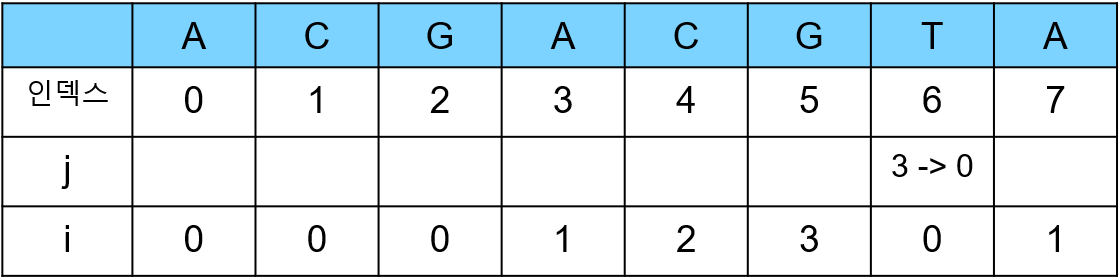
**<KMP 작동 원리>**

**1. ShortRead 파일에서 한 줄을 읽어와서 문자열에 저장한다.**

ShortRead = “ACGACGTA” 라고 가정한다.

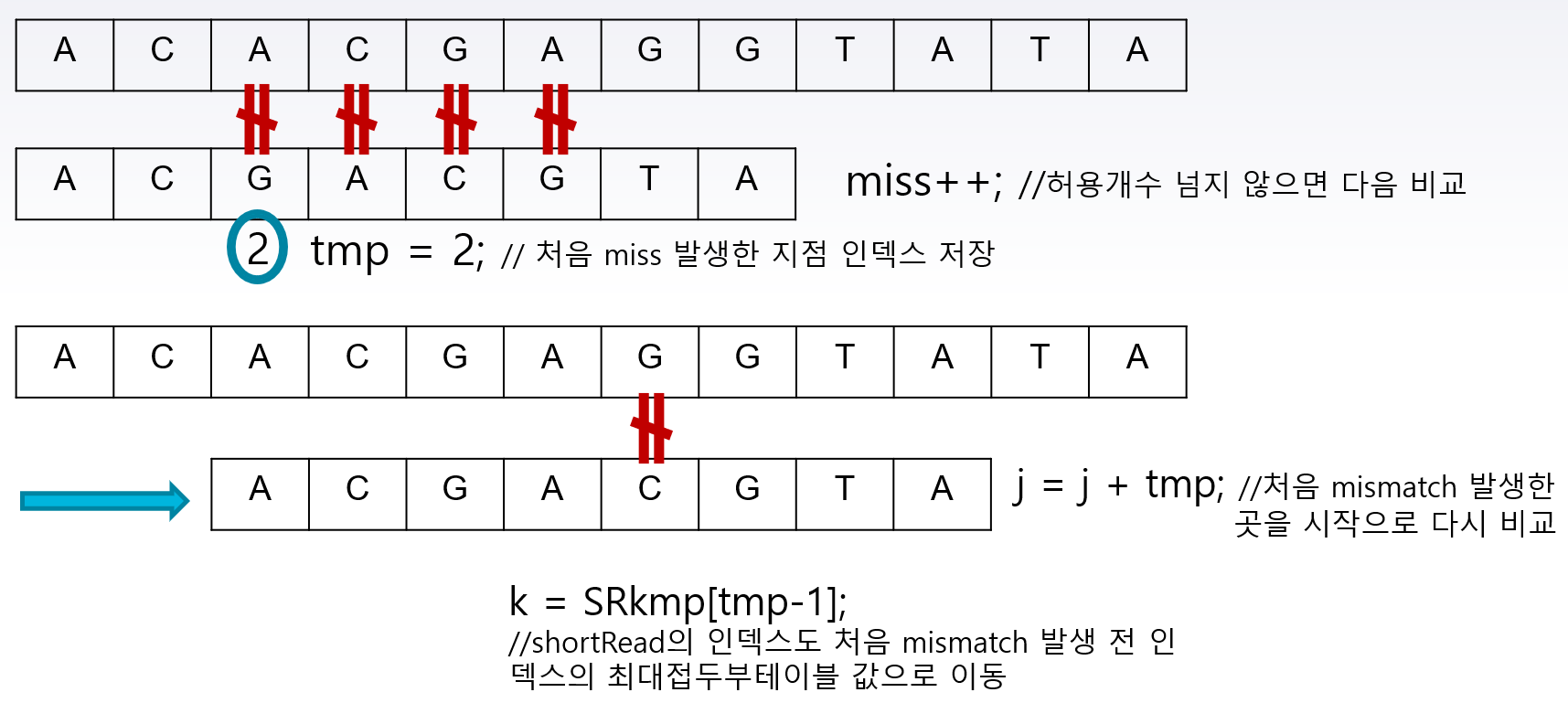
**2. 저장한 문자열의 최대접두부 테이블을 생성한다.**



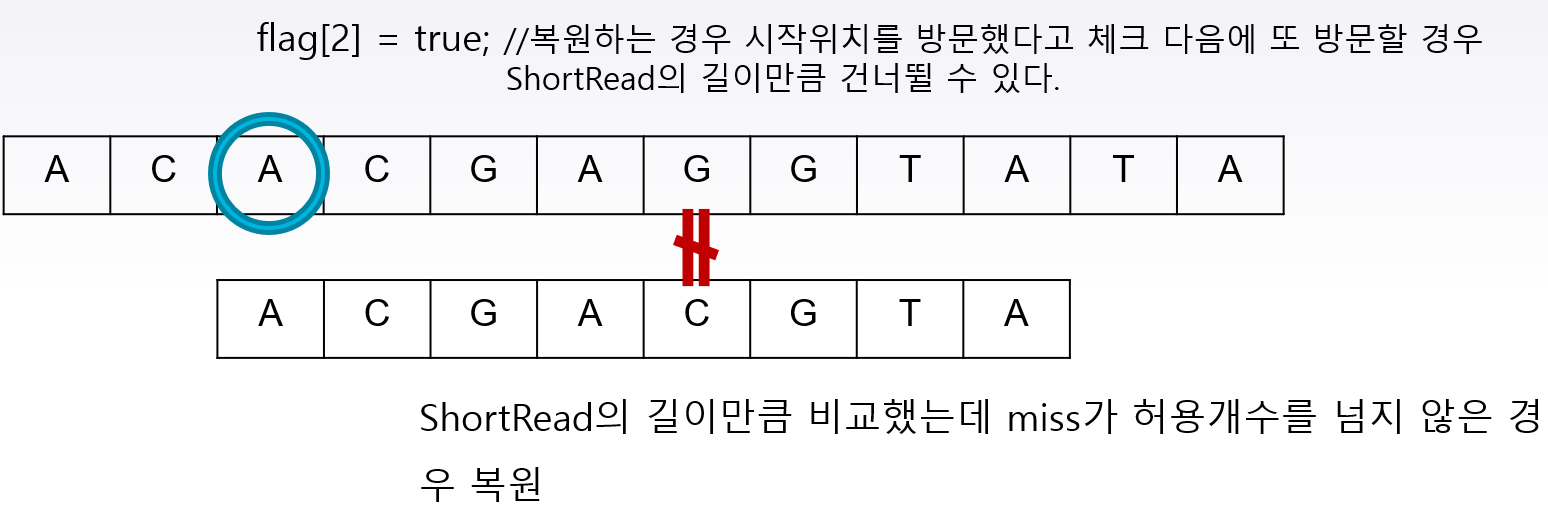
**3. 반복문을 통해 ShortRead와 Reference DNA sequence를 비교한다.**

(1) 비교를 시작하는 지점이 방문한 인덱스라면 이미 복원을 한 지점이므로 ShortRead의 길이만큼 건너뛴다.

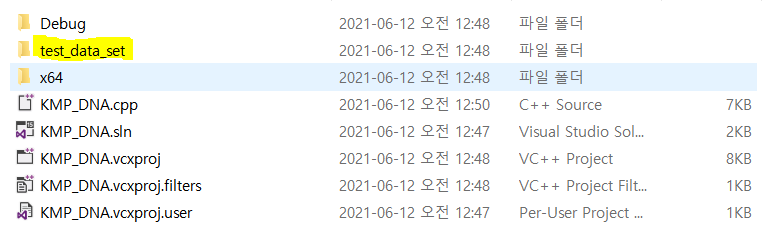
(2) ShortRead와 Reference DNA sequence가 일치하지 않을 때 0으로 초기화 한 정수형 변수 miss의 수를 1 증가시킨다. miss가 처음으로 발생하는 지점의 위치(인덱스)를 정수형 변수 tmp에 저장한다. miss의 수가 초기에 설정한 miss match 허용개수 이하일 경우 반복문을 계속 진행시키고 초과한다면 어느 지점부터 다시 비교를 시작할지 인덱스를 조정하고 반복문을 탈출한다. 여기서 최대 접두부 테이블과 처음으로 miss가 발생한 위치를 저장한 tmp를 사용한다.



(3) miss가 miss match 허용개수 이하인 상태로 ShortRead의 길이만큼 비교를 완료했다면 그 부분의 문자열을 복원해주고 복원 시작 위치를 방문한 곳으로 체크한다.



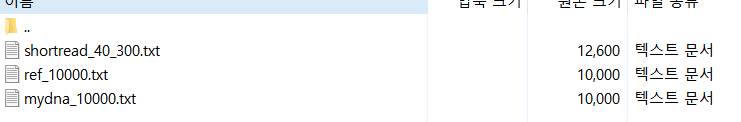
**<시연에 사용할 KMP 프로그램 설명>**



위의 그림처럼 프로젝트 파일 안에 test\_data\_set 파일을 저장합니다.

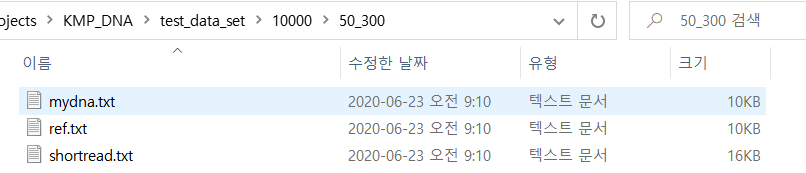
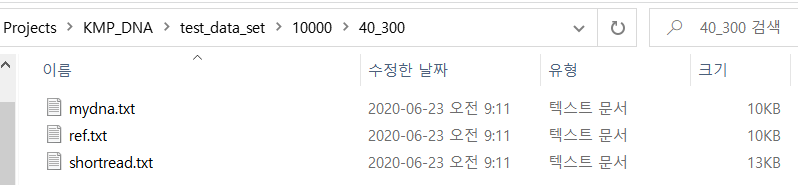
test\_data\_set의 10000파일의 .txt파일들의 이름에만 차이가 있어서 동일하게 바꿔줬습니다.

**<기존>**





**<변경 후>**



ifstream ifs1("test\_data\_set\\" + to\_string(DNA\_SIZE) + "\\" + to\_string(Length) + "\_" + to\_string(Num) + "\\shortread.txt");

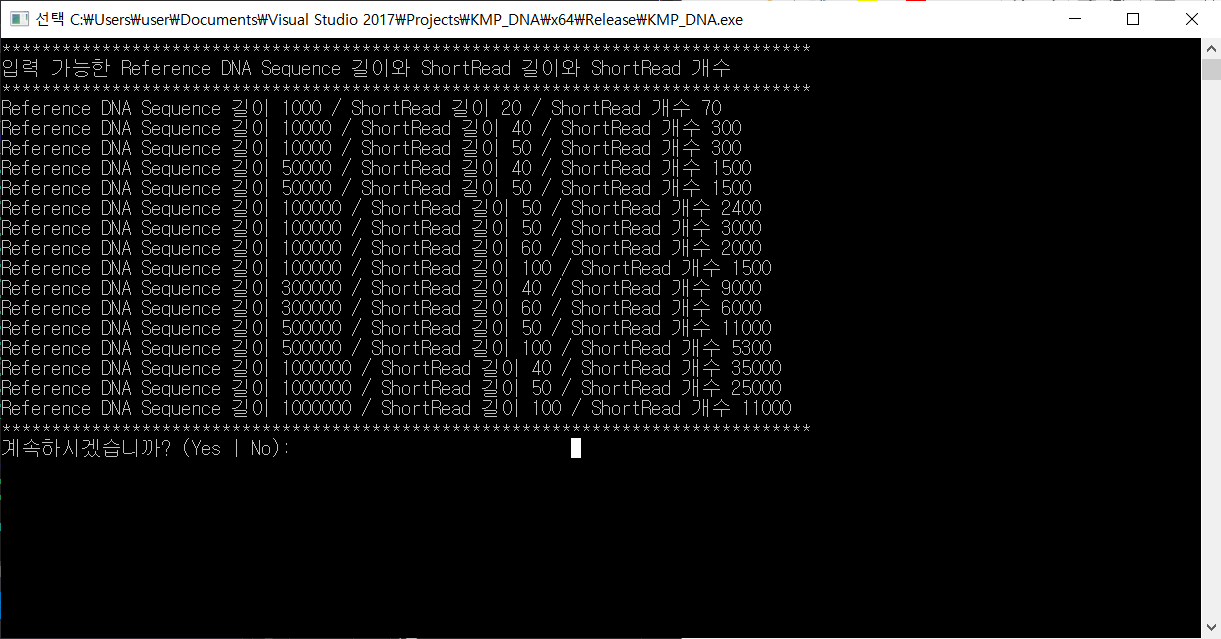
ifstream ifs2("test\_data\_set\\" + to\_string(DNA\_SIZE) + "\\" + to\_string(Length) + "\_" + to\_string(Num) + "\\ref.txt");

ifstream ifs3("test\_data\_set\\" + to\_string(DNA\_SIZE) + "\\" + to\_string(Length) + "\_" + to\_string(Num) + "\\mydna.txt");

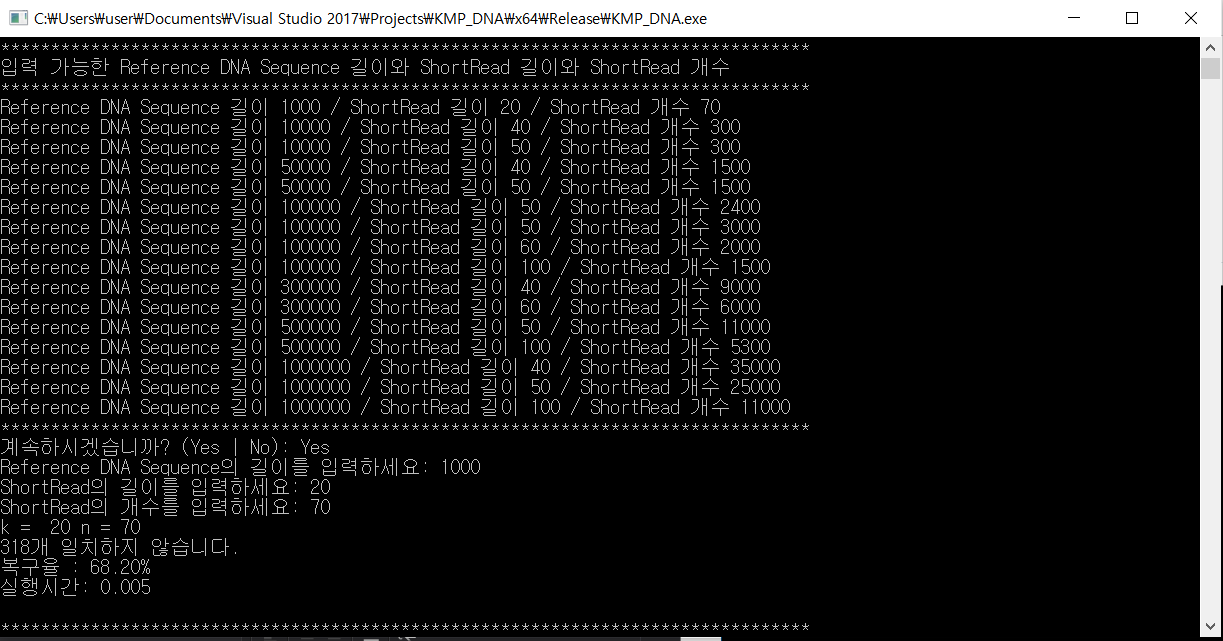
Reference DNA Sequence의 길이를 DNA\_SIZE로 하고 ShortRead의 길이를 Length로 ShortRead의 개수를 Num으로 입력받고 위와 같은 코드를 통해서 각각 shortread.txt / ref.txt / mydna.txt 파을 열고 문자를 읽어와서 KMP 알고리즘으로 mydna 시퀀스 복원을 시작한다.

**<프로그램 사용방법>**

프로그램을 실행시키면 아래와 같은 화면이 출력된다.



계속하려면 Yes(No, nO, no를 제외한 아무런 키를 입력해도 상관없음)를 입력해준다. 그리고 위의 표에 나온 것에서 골라서 길이와 개수를 입력해준다.



Reference DNA Sequence의 길이를 1000으로 입력하고 ShortRead의 길이를 20 개수를 70으로 입력하면 해당 폴더로 가서 파일을 열고 복원한다. 그리고 해당 결과를 출력하고 반복문을 통해서 다시 위와 같은 화면이 출력된다.

프로그램을 멈추기 위해서는 No (No, nO, no 모두 가능)를 입력해준다. 만약 출력화면 위에 나와있는 길이와 개수가 아닌 다른 길이나 개수를 입력하면 존재하지 않는 폴더나 파일에 접근하게 되어 에러가 발생한다. 그래서 파일을 찾을 수 없음!! 이라는 문구와 함께 프로그램은 종료한다.

