|  |
| --- |
| C++프로그래밍및실습  DNA 해독기  **진척 보고서 #2** |

|  |  |
| --- | --- |
|  | 제출일자: 2023.12.14  제출자명: 백성준  제출자학번: 175989 |

1. 프로젝트 목표
   1. 배경 및 필요성

“Chat gpt"와 같은 대화형 인공지능의 출현으로 세계의 기술적 환경은 크게 변화하였다. 평범한 사람들도 이제는 그 변화를 피부로 느끼고 있다. 당장 이러한 대화형 인공지능을 사용해본 사람들은 그 용이성과 위기감을 느낄 수 있다. 머지않아 대부분의 단순 노동은 대체될 것이며, 가장 인간적이라 생각한 부분도 대체될 것이다. 그리고 실제로 지금도 대체가 일어나고 있다.

이러한 변화 속에서 인공지능 또는 자동화 기술을 다양한 학문과 연계한다면 큰 경쟁력을 획득할 수 있을 것이다. 본 프로젝트에서는 생물학에서의 자동화를 목표로 한다.

* 1. 프로젝트 목표

이 프로젝트에서는 생명의 가장 기초적이지만 근본적인 DNA 염기서열을 자동으로 처리하는 프로그램을 다루고자 한다. DNA 염기서열에서 ORF(Open Reading Frame)을 찾아내고 CDS(Coding sequence)를 최종적으로 얻어내는 것을 목표로 한다. 또한, 이 CDS를 단백질로 변환하는 작업까지 수행한다. 이 때, 사람의 DNA라 가정하며 가장 자주 발견되는 consensus sequence를 따른다고 가정한다. 또한, codon table은 보편적인 table을 사용한다.

* 1. 차별점

기존의 염기서열 정보를 단백질 서열로 변환하는 프로그램에는 단순 변환만 존재하였다. 여기에서 더 나아가 DNA에서 얻어낸 단백질의 정보가 저장된 ORF(Open Reading Frame)에서 intron을 제거해낼 수 있다. 또한, 얻어낸 서열 정보의 분석을 진행한다.

* 1. 배경지식

사람의 유전자는 A,T,G,C의 염기를 가지고 있다. 이 염기가 3 개 모여서 단백질 정보를 이루는 것을 codon이라 하고, codon은 20 개 단백질 정보를 담고 있다. 각 codon에 하나의 단백질이 상응하며 단백질의 시작이 되는 start codon과 stop codon이 각각 존재한다. Start codon에서 stop codon 까지를 ORF(Open Reading Frame)라 한다.

DNA는 RNA로 변환되어 splicing이라는 intron이 제거되는 과정을 거친다. Intron이 제거된 RNA가 ribosome에서 단백질로 변환된다. 이때의 RNA서열에 상응하는 DNA를 CDS(Coding Sequence)라 한다.

1. 기능 계획
2. 사용자의 이름과 염기서열 정보, 찾을 ORF 범위 저장
3. 사용자의 이름을 입력
4. 염기서열의 정보를 입력
5. 찾을 ORF의 크기를 입력
6. 염기서열의 정보를 분석
7. 염기서열에서 ORF를 저장

-6 가지의 가능한 reading frame을 저장한다.

-각 서열에서 ATG로 시작하고, TGA, TAA 또는 TAG로 끝나는 서열을 저장한다.

1. 얻어낸 ORF에서 intron을 제거

- [GTA][AGT]로 시작하며 [TTT][TTT][TT ][CAG]로 끝나는 서열을 제거한다.

1. Intron이 제거된 CDS를 단백질 형태의 정보로 변환

-codon table에 맞게 염기서열을 단백질 형태로 변환한다. 즉, 각 codon에 맞는 amino acid로 변환한다.

1. Kozak sequence score를 계산

-염기서열에서 +4 위치에서 C 또는 G가 나오면 점수를 1 올리고, -2 위치에 A 또는 G가 존재하면 마찬가지로 1점 올리고, -3 에서 A 또는 G가 존재하면 +3점 증가한다. 이 점수를 합산하여 Kozak strength를 측정하여 출력시킨다.

**3. 진척사항**

**1) 기능 구현**

1. 얻어낸 ORF에서 intron을 제거

- 입출력

Orf user\_seq1, Orf user\_seq2, Orf user\_seq3, Orf user\_seq4, Orf user\_seq5, Orf user\_seq6 : 유저의 orf를 처리하기 위한 클래스 변수

void Orf::IntronFinder() : user\_seq1~6에 존재하는 intron\_removed 객체 변수에 ORF에서 인트론 부분을 수정하여 저장시키는 함수.

vector<vector<string>> Orf::intron\_removed : Orf의 객체 변수로 기존의 ORF에 인트론을 따로 표시하여 저장시켜놓기 위한 변수.

- 설명

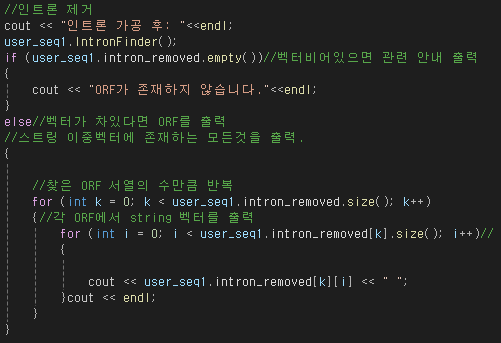
찾은 ORF들을 하나씩 가져와서 인트론에 해당하는 부분을 찾고 그 부분을 “intron” string으로 바꿔서 string 이중 벡터에 저장한다. 또한, 이를 출력하여 그 결과를 확인한다.

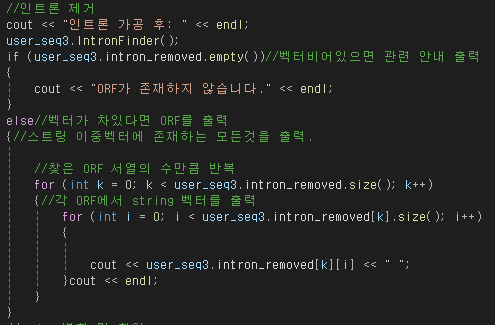
만약, 가공할 ORF가 없다면 오류메시지 출력.

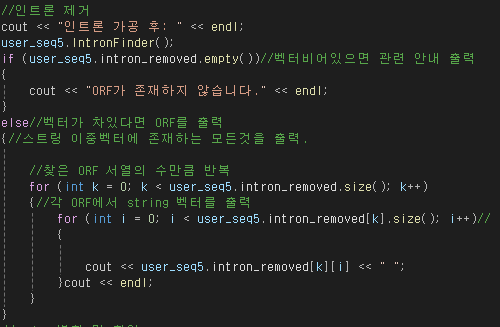
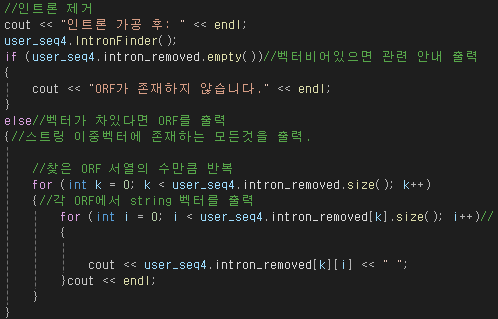
- 적용된 배운 내용

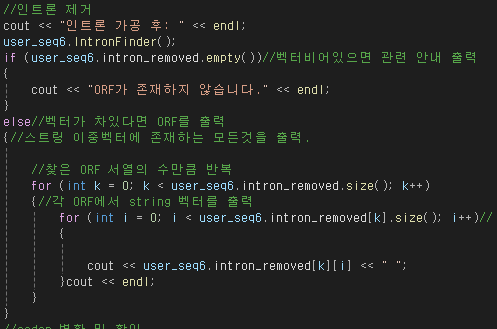
클래스, cin/cout 입출력문, vector의 멤버함수(size(),empty()), 이중 for문, if else문, 배열, 스트링, 벡터, 증감연산자, 비교연산자

- 코드 스크린샷









1. Intron이 제거된 CDS를 단백질 형태의 정보로 변환

- 입출력

Orf user\_seq1, Orf user\_seq2, Orf user\_seq3, Orf user\_seq4, Orf user\_seq5, Orf user\_seq6 : 유저의 orf를 저장하기 위한 클래스 변수

void Orf::CodonDecipher() : 클래스 변수내의 인트론이 제거된 ORF가 단백질 정보를 갖도록 클래스 변수를 변환하는 함수.

vector<vector<string>> ORF::protein : 스트링 이중 벡터로, 단백질 정보로 변환된 ORF를 저장하기 위한 클래스 변수.

- 설명

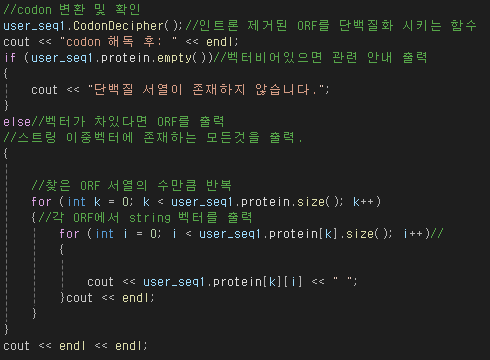
앞에서 얻어낸 인트론이 제거된 ORF인 CDS를 단백질 형태의 정보를 갖도록 변환한다. 또한, 이러한 변환 결과를 출력하여 확인이 가능하게 한다.

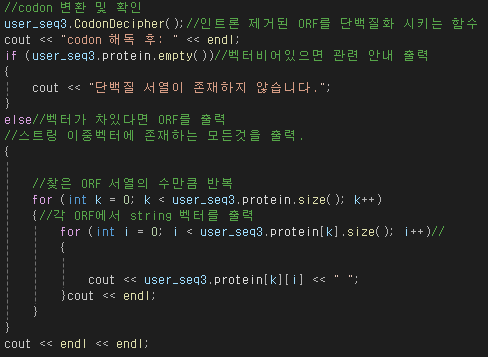
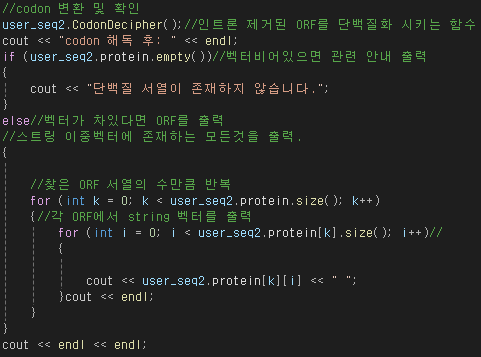
만약, 가공할 ORF가 없다면 오류메시지 출력.

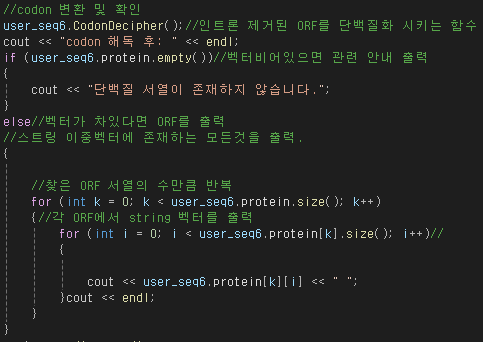
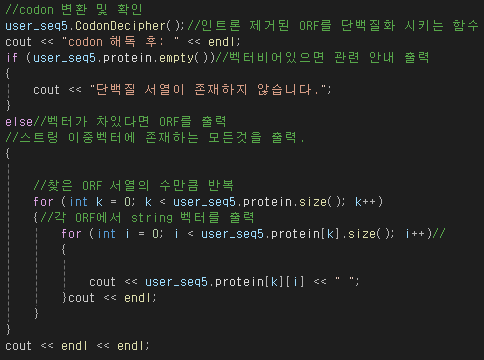
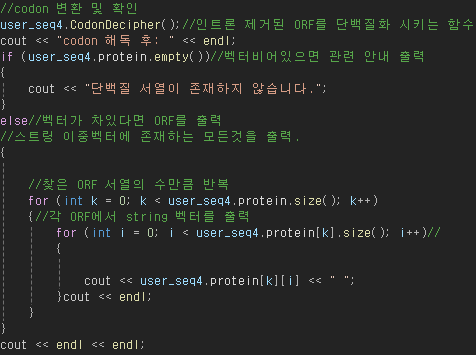
- 적용된 배운 내용

클래스, cin/cout 입출력문, vector의 멤버함수(size(),empty()), 이중 for문, if else문, 배열, 스트링, 벡터, 증감연산자, 비교연산자

- 코드 스크린샷







**1-1) 함수 구현**

1. void Orf::IntronFinder()

- 입출력

vector<vector<string>> Orf::complete\_orf : 인트론을 찾아서 모아놓은 스트링 이중 벡터. 이것을 가공하기 위해 가져옴.

vector<vector<string>> Orf::intron\_removed : 위의 객체 변수 complete\_orf를 넣고, 여기서 인트론을 찾아서 인트론을 표시하여 저장해 놓기 위한 객체 변수.

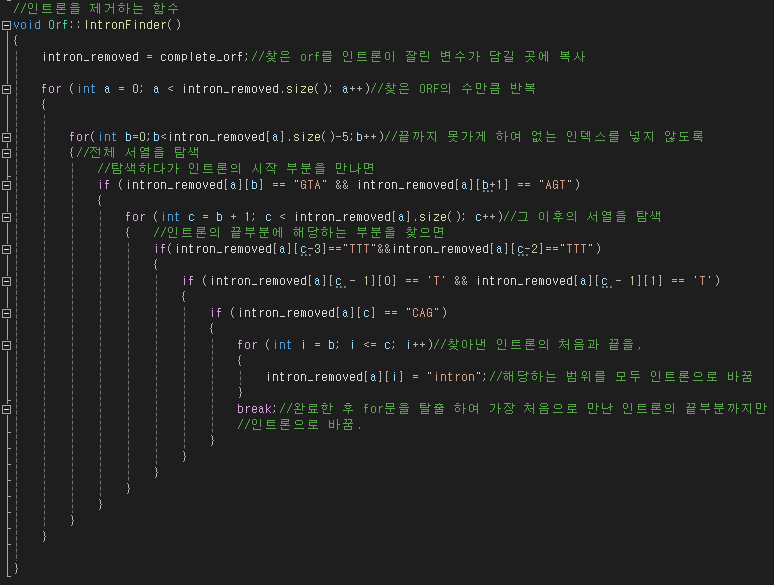
- 설명

전에 찾은 ORF를 가져와서, 새로운 변수에 담고 이 새로운 변수에 가공을 한다. 인트론의 시작부분에 해당하는 부분과 끝 부분에 해당하는 부분을 찾으면, 해당 범위는 모두 string 형의 “intron”값을 갖도록 변환한다.

- 적용된 배운 내용

클래스, 이중 for문, vector의 멤버 함수(size()), 스트링, 비교연산자, 논리연산자, 배열, break, 벡터, 증감연산자

- 코드 스크린샷



1. void Orf::CodonDecipher()

- 입출력

vector<vector<string>> Orf::intron\_removed : 인트론이 표시된 ORF를 저장해 놓은 객체 변수. 이것을 가져와 단백질 정보로 변환하기 위함.

vector<vector<string>> Orf::protein : 가져온 인트론이 표시된 ORF를 단백질 정보로 변환하여 저장해 놓기 위한 객체 변수.

string pro\_cpy : 스트링 이중 벡터의 각 스트링 값을 받아오는 지역변수. 이 값을 이용해서 저장된 값을 확인하고 혹시나 생길 기존 서열의 변환을 방지 한다.

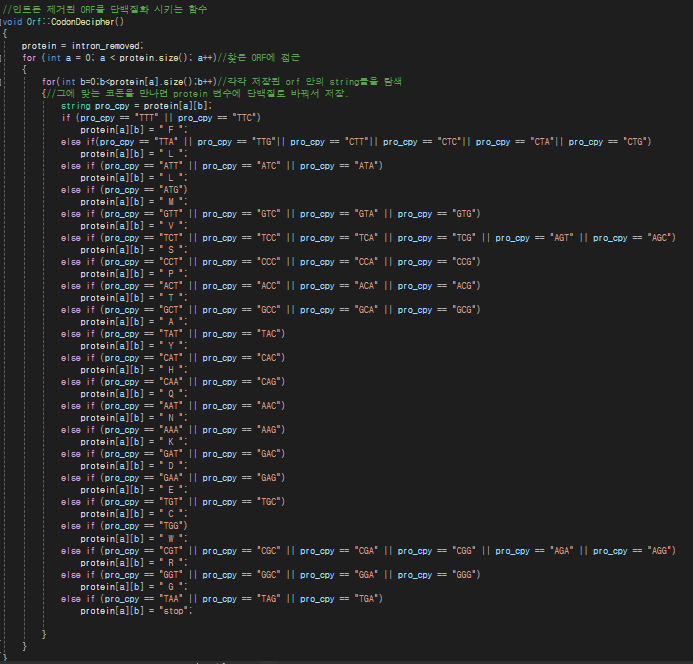
- 설명

인트론이 표시된 ORF의 값들을 가져와서 새로운 변수에 옮겨 넣고, 이 변수에 각 단백질 codon에 해당하는 서열이 나오면 이를 단백질 형태의 정보로 변환하여 저장한다.

- 적용된 배운 내용

이중 for문, 벡터, 비교연산자, 논리연산자, 배열, if문, 벡터의 객체함수(size()), 증감연산자

- 코드 스크린샷



**2) 테스트 결과**

**(1)** 얻어낸 ORF에서 intron을 제거

- 설명

ORF가 존재하지 않는 경우.

- 테스트 결과 스크린샷



- 설명

인트론이 한 ORF에 하나만 존재하는 경우.

- 테스트 결과 스크린샷



- 설명

인트론이 한 ORF에 여러 개 존재하는 경우.

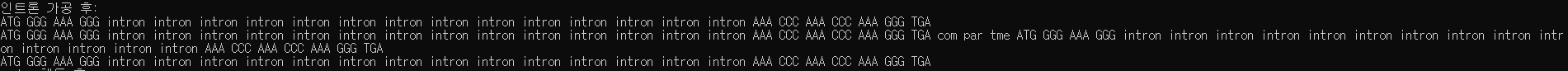
- 테스트 결과 스크린샷



- 설명

하나의 reading frame에서 여러 개의 ORF를 찾아서, 다수의 가공할 ORF가 존재하는 경우.

- 테스트 결과 스크린샷



- 설명

ORF는 있지만 그 서열에서 인트론을 찾지 못한 경우.

- 테스트 결과 스크린샷



1. Intron이 제거된 CDS를 단백질 형태의 정보로 변환

- 설명

해독할 ORF가 존재하지 않는 경우.

- 테스트 결과 스크린샷



- 설명

순서대로 가능한 코돈을 모두 출력 입력한 경우.

- 테스트 결과 스크린샷



- 설명

해독할 ORF가 하나인 경우.

- 테스트 결과 스크린샷



- 설명

해독할 ORF가 여러 개 존재하는 경우.

- 테스트 결과 스크린샷



**2-1) 함수 테스트 결과**

1. void Orf::IntronFinder()

- 설명

인트론이 한 서열에 하나만 존재하는 경우.

- 테스트 결과 스크린샷



- 설명

인트론이 한 서열에 여러 개 존재하는 경우.

- 테스트 결과 스크린샷



- 설명

그 서열에서 인트론을 찾지 못한 경우.

- 테스트 결과 스크린샷



1. void Orf::CodonDecipher()

- 설명

순서대로 가능한 코돈을 모두 입력한 경우.

- 테스트 결과 스크린샷



- 설명

해독할 ORF가 하나인 경우.

- 테스트 결과 스크린샷



- 설명

해독할 ORF가 여러 개 존재하는 경우.

- 테스트 결과 스크린샷



**4. 계획 대비 변경 사항**

1. 사용자의 이름과 염기서열 정보, 찾을 ORF 범위 저장

-이전

1. 사용자의 이름을 입력
2. 염기서열의 정보를 입력
3. 찾을 ORF의 크기를 입력

-이후

1. 사용자의 이름을 입력
2. 염기서열의 정보를 파일로 입력
3. 찾을 ORF의 크기를 입력

-사유

크기가 방대한 염기서열의 특성상 파일로 입력하는 것이 더 합리적이다.

1. 염기서열의 정보를 분석

-이전

1. 얻어낸 ORF에서 intron을 제거

- [GTA][AG ] 또는 [GTG][AG ]로 시작하며 [CAG]로 끝나는 서열을 제거한다. 이 때, 중간에 A가 하나 이상 존재해야 한다.

-이후

1. 얻어낸 ORF에서 intron을 제거

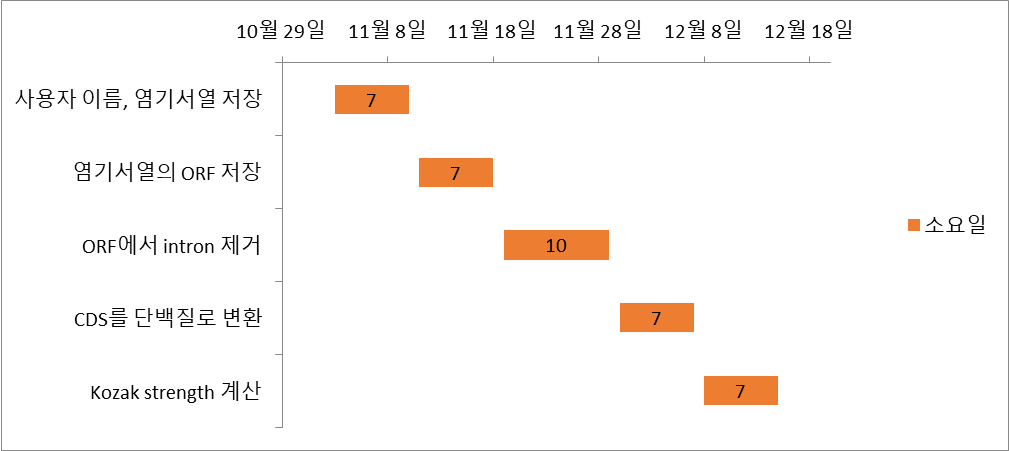
- - [GTA][AGT]로 시작하며 [TTT][TTT][TT ][CAG]로 끝나는 서열을 제거한다.

-사유

중간에 A 염기서열이 없을 확률은 매우 적고 없어도 확인이 가능하다. 또한, 인트론의 서열은 확실히 정해진 것이 아니기 때문에 많은 경우의 수가 존재하고 이것을 유추하여 완벽히 찾아낸다는 것은 현재의 과학 수준에서는 불가능하다. 현재 이에 대한 대안으로 기존의 서열과 비교해서 찾는 방법이 존재한다.(BLAST 등) 따라서, 이 프로젝트에서는 가장 많이 발견되는 consensus 서열과 일치하는 확실한 경우에만 인트론을 제거하기 위해 변경하였다.

**5. 프로젝트 일정**

(진행한 작업과 진행 중인 작업 등을 표기)



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 일정명 | 시작일 | 소요일 | 마감일 |  |
| 사용자 이름, 염기서열 저장 | 11월 3일 | 7 | 11월 10일 | 완료 |
| 염기서열의 ORF 저장 | 11월 11일 | 7 | 11월 18일 | 완료 |
| ORF에서 intron 제거 | 11월 19일 | 10 | 11월 29일 | 완료 |
| CDS를 단백질로 변환 | 11월 30일 | 7 | 12월 7일 | 완 |
| Kozak strength 계산 | 12월 8일 | 7 | 12월 15일 |  |