1. **Henkilötiedot**
   Proteiinisynteesi

   Maria Ushakova, 829825
   Bioinformaatioteknologiat, vuosikurssi: 2
   24.02.2022

2. **Yleiskuvaus ja vaikeustaso**

   Tee ohjelma, joka mallintaa proteiinisynteesin päätapahtumia, eli RNA:n lukemista DNA:n toisesta juovasta ja peptidiketjun muodostumista RNA:n emästen mukaan.
   Eri emästen pitää olla ohjelmassa erotettavissa toisistaan (*U* ja *T* saavat olla samannäköisiä) ja emäsparien (*A* – *U*/*T* ja *C* – *G*) pitää näyttää kuuluvan yhteen, toisin kuin väärien yhdistelmien.
   DNA-ketjun lukeminen RNA:ksi pitää aloittaa vasta DNA-ketjussa esiintyneen TATA-box:n, eli sekvenssin *TATA*, jälkeen (tämän ei tarvitse näkyä simulaatiossa). RNA:n lukemista ei kuitenkaan kuulu aloittaa heti TATA-box:n jälkeen vaan TATA-box:n jälkeen tulevasta aloituskodonista DNA:ssa *TAC* (RNA:ssa *AUG* ja peptidinä metioniini) jälkeen.
   Proteiinin peptidiketju kuuluu muodostaa RNA-ketjun mukaisesti, niin että ketjun seuraava aminohappo valitaan aina RNA-ketjun seuraavan kolmen peräkkäisen emäksen mukaan. Taulukon aminohappojen kodoneista löydät täältä. Lopeta peptidiketju lopetusaminohapon (3 erilaista) kohdalla.
   Simulatiossa pitää olla DNA:n toinen juoste, jota luetaan, lähetti-RNA ja peptidiketju. Ohjelmassa ei tarvitse olla rna-polymeraasia, siirtäjä-rna:ta, ribosomia, tms. Vaikean työn yksi kriteeri on erotella RNA:n eksonit ja intronit ja poistaa intronit, eli silmukointia ei tarvitse huomioida keskivaikeassa lainkaan.
   Ohjelman on tarkoitus lukea DNA-sekvenssi tiedostosta, eli ohjelman kuuluu toimia useammalla sekvenssillä eikä vain yhdellä. Erilaisia DNA-sekvenssejä saat tehtyä esimerkiksi random DNA sequence generaattorilla (suositeltu pituus 500-1000).

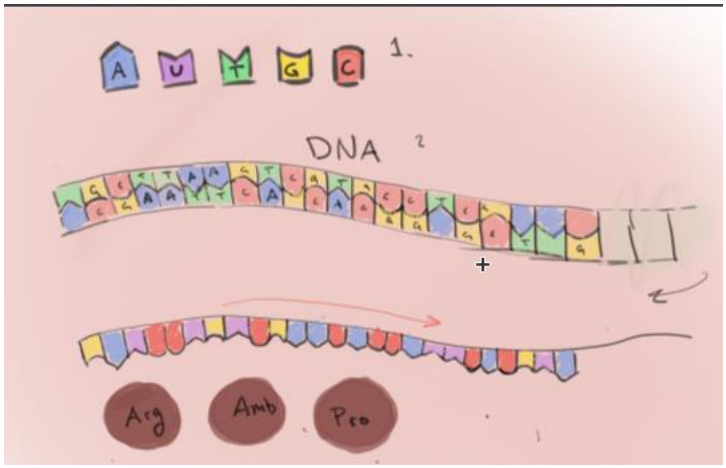
- DNA-> RNA lukemisessa mahdollisuus virheille, DNA:n lukemisessa tapahtuu virhe noin 10^-4 todennäköisyydellä, mutta saat itse valita virhetodennäköisyyden
  - silmukointi, eli intronien poistaminen DNA:sta
  - intronit voi erotella eksoneista, esimerkiksi että introni alkaa aina GU:lla ja loppuu AG:hen (tämä on yleistä introneille)
  - introneita ei tarvitse kirjaimellisesti silmukoida pois, riittää kun ne poistetaan sekvenssistä, niin että poiston huomaa ohjelman suorituksessa (intronit voi esimerkiksi tehdä eri värillä kuin eksonit, ja poistaa yksitellen sekvenssistä)

   vaikeustaso:

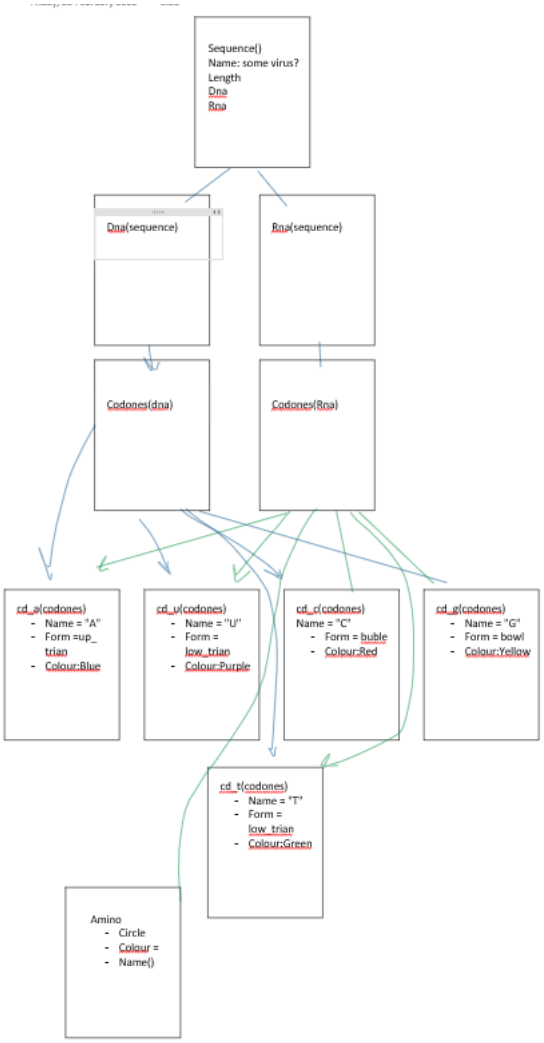
3. **Käyttötapauskuvaus ja käyttöliittymän luonnos**

Ohjelma kommunikoi käyttäjän kanssa graaffisesti. Alussa tulee esiin näkymä, missä näkyy kiva kuva proteinisynteesistä ja missä käyttäjältä pyydetään tiedoston nimi, mitä käytetään projektissa. Voisi myös lisätä allustavasti liitetyt dna-sequencit joistain kuuluisita viruksen dna:sta tai muuta vastaavaa.

Sen jälkeen kun käyttäjä on kertonut tiedoston nimen, ohjelma aloittaa koko projektin. Käyttäjälle näytetään DNA-ketjua, missä jokainen kodoni on piirretty omassa muodossa, omalla väirillä ja kirjaimella. Sen jälkeen piirretään vastaavalla tavalla RNA- ketju. En ole vielä varma otanko haastavamman version projektissa, missä teen silmukoinnin, mutta jos teen, käyttäjälle näytetään myös se. Sitten muodostetaan ja piirretään kodoneista muodostettuja aminohappo-kuplia.



4. Ohjelman rakennesuunnitelma



Koen, että juuri semmoiset luokat sopivat projektiin, koska melkein jokaisen osan täytyy piirtää erikseen. On myös paljon samantyyppisiä osia, kuten kodonit, tai ketjut, joille on tehty parent-class. Koko projektissa tärkeimpänä osana ovat ketju, joka koostuu kodoneista (perinnöllisyys näkyy selkeästi kartalla Sequence -> dna/rna -> Codone(rna)/Codone(dna)).

5. Tietorakenteet

Projektissani tulen käyttämään taulukon aminohappojen RNA kodoneista:

```
table = {
        'ATA':'I', 'ATC':'I', 'ATT':'I', 'ATG':'M',
        'ACA':'T', 'ACC':'T', 'ACG':'T', 'ACT':'T',
        'AAC':'N', 'AAT':'N', 'AAA':'K', 'AAG':'K',
        'AGC':'S', 'AGT':'S', 'AGA':'R', 'AGG':'R',
        'CTA':'L', 'CTC':'L', 'CTG':'L', 'CTT':'L',
        'CCA':'P', 'CCC':'P', 'CCG':'P', 'CCT':'P',
        'CAC':'H', 'CAT':'H', 'CAA':'Q', 'CAG':'Q',
        'CGA':'R', 'CGC':'R', 'CGG':'R', 'CGT':'R',
```

```
        'GTA':'V', 'GTC':'V', 'GTG':'V', 'GTT':'V',
        'GCA':'A', 'GCC':'A', 'GCG':'A', 'GCT':'A',
        'GAC':'D', 'GAT':'D', 'GAA':'E', 'GAG':'E',
        'GGA':'G', 'GGC':'G', 'GGG':'G', 'GGT':'G',
        'TCA':'S', 'TCC':'S', 'TCG':'S', 'TCT':'S',
        'TTC':'F', 'TTT':'F', 'TTA':'L', 'TTG':'L',
        'TAC':'Y', 'TAT':'Y', 'TAA':'_', 'TAG':'_',
        'TGC':'C', 'TGT':'C', 'TGA':'_', 'TGG':'W',
    }
```

6. Tiedostot ja tiedostoformaatit

Projektissani tietoa käsitellään tekstitiedostona, jota ohjelma lukee ja lajittelee luokkaan.

7. Algoritmit

Ohjelma lukee tekstiä tiedostosta ja lataa sen, sen jälkeen lajittelee luokkaan sequence -> DNA. Sitten graaffissen kirjaston kautta printataan jokaisen kodonin erikseen (sitä varten luokat kodoneille) Dna:sta me sijoitetaan T-kodonin sijaan U-kodonin . Sen jälkeen piirtäminen tosituu, mutta jo Rna-kodonketjun muodossa.

8. Testaussuunnitelma

Testausohjelmiin kuuluu semmoiset testit kuten:

tiedoston löytäminen

Tarkistus, että tarvittava ketju sisällyttää tarvittava määrä kodoneja

Trakistus, että RNA ketju muodostuu tarvittavasta määrästä tripletteja

Tarkistus, että oliot ovat oikein luokiteltuja

9. Kirjastot ja muut työkalut

Käytän projektissani PyQt kirjastoa.

10. Aikataulu

Ensimmäisen välipalautuksen pitäisi olla jo kehitettynä oliot ja niihin liittyvät testit: 10-20h
Toisen palautukseen pitäisi olla kehittynyt pää ohjelmat ja niiden testi ohjelmat: 15-20h
Kolmanteen palautukseen pitäisi olla tehty main ja kokonaisuus koodi:15-25h
Viimeisenä (4) – graaffinen osuus: 10-15h

11. Kirjallisuusviitteet ja linkit

A+
https://plus.cs.aalto.fi/y2/2022/project_topics/topics_simulaatiot_212proteiini/
Youtube video proteiinisynteesistä:
https://www.youtube.com/watch?v=oefAI2x2CQM&ab_channel=AmoebaSisters
https://www.youtube.com/watch?v=fBetNoFd8jo&t=471s&ab_channel=Opetus.tv
esimerkki samantyyppisestä tehtävästä:
https://www.geeksforgeeks.org/dna-protein-python-3/

12. Liitteet

Lisäksi suunnitelmassa saa olla liitteitä, aiheesta riippuen.