Предлог проект
07.04.2021

ВОВЕД ВО БИОИНФОРМАТИКА

Опис на предлог проект

Предвидување на мутација на глико-протеинот на SARS-Cov2 вирусот

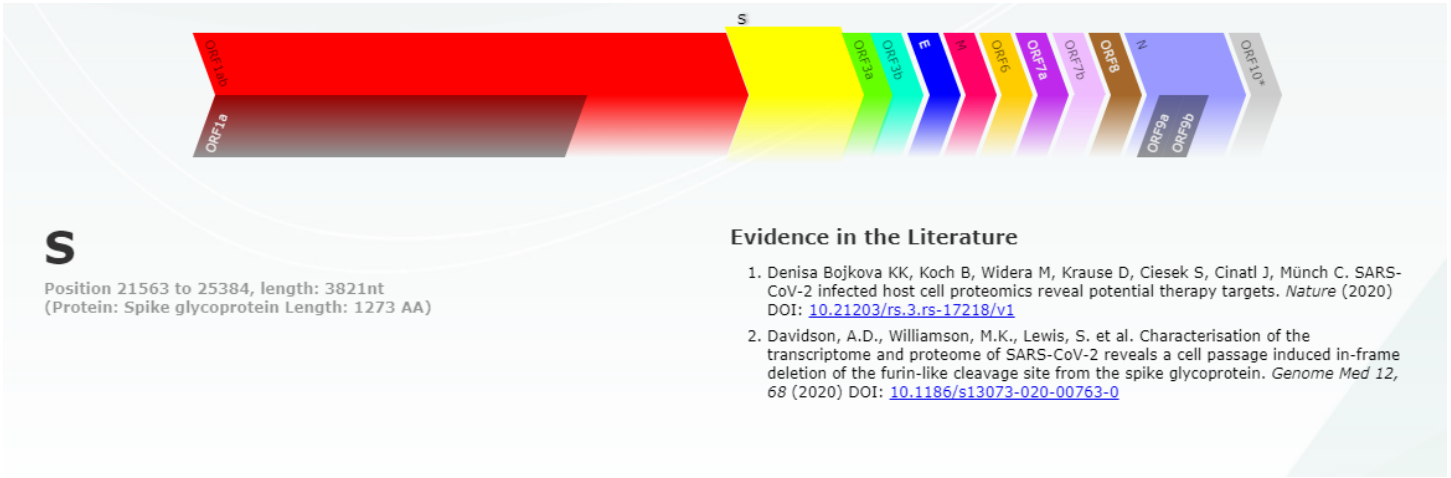
Професор*:* д-р Невена Ацковска
Студент*:* Кирил Зеленковски, 161141

0. ИДЕЈА

Предвидување на мутација на кодниот регион т.е. ген S кај **SARS-Cov2** вирусот. Ова е протеинскиот глико-протеин познат како **spike protein.**

На влез се чита РНА секвенца (од тој регион: 21563-25384) на излез се добива:

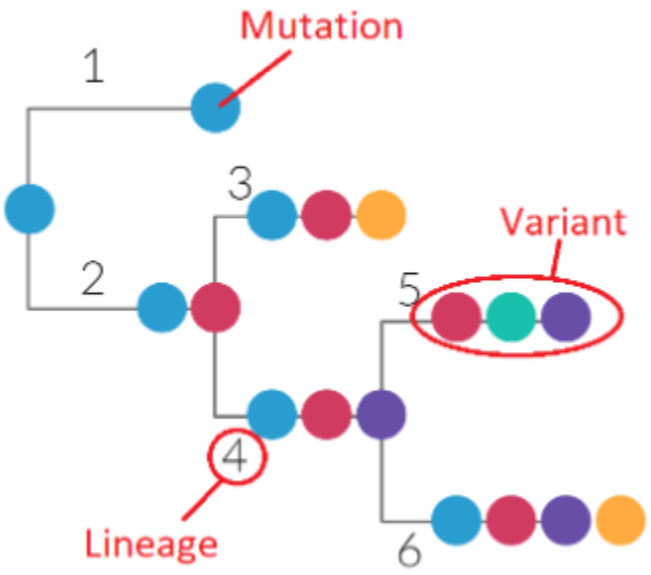
- Една идна мутација (прв чекор)
- Повеќе идни мутации (втор чекор)



Фиг1. Поделба на кодни региони кај РНА на SARS-Cov2 ((WIV04)

Истражувањето премногу зависи од множеството за учење и тренирање. Идејата е да ги соберам сите можни мутации во едно множество и со овој груба идеја да се предвидуваат идни врз основа на промени или бришења на бази во претходно мутации. Енормно значење има time stamp на овие мутации.

Сите податоци би биле собирани од **GISAID**.



Фиг 2. Разлика меѓу мутација, варијанта и lineage

1. *ДЕТАЛИ (или продолжување на иницијалната идеја)*

Класификација на мутации

SARS-Cov2 е **РНА** базиран вирус, што се многу различни од **ДНА** вируси во контекст на повисоки рати на мутирање со што имаат повисока стапка на адаптивност

Откако ќе се предвидат мутациите може да се направи и класификација врз база на каква мутација настанала. Има неколку видови на мутации кои настануваат на **РНА** секвенцата (кои што би биле класата за предвидувње откако ќе ги добиеме):

- **point mutation** (**base substitutions**): промени од мал размер кои настануваат на РНА секвенцата за 1 до 2 нуклеотида еден ваков вид мутација е **нуклеотидна замена**
- **frame shift** (**deletions** или **insertions**): се однесуваат на додавање (insertion) или бришење (deletion) на нуклеотиди во РНА секвенцата.

Друг проблем кои истотака може да се истражи откако ќе се добијат, е какво е нивното влијание врз основа на нивното значење:

- **silent mutation:** промена на кодон со што резултатната амино секвенца останува непроменета
- **missense mutation:** промена која влијае на резултатна амино секвенца
- **nonsense mutation:** промена на кодон која продуцира стоп кодон кој потоа во генската транслација предизвикува нефункционални читања

Rough set theory за предвидување на РНА мутации

Техника што предвидува потенцијални замени што може да настануваат во примарната **РНА** секвенца [1]. Во секоја итерација на учење:

- влез: РНА секвенца од генерација (мутација)
- излез: РНА секвенца од следна генерација (мутација)

Features на влезот се сите нуклеотиди во **РНА** секвенцата и истите кореспондираат со features на излезот. Моделот го храниме со пораменти **РНА** секвенци од последователни генерации (**GISAID**, outbreak.info) од истиот вирус (истата РНА) што се секвенционирани во слична околина (нас ни треба човек)
Крајната цел е да се предвиди следната РНА секвенца зимајќи ги во предвид сите претходни РНА секвенци.

Евалуација на моделите

Се собираат вакви мутации и предвидените РНА секвенци се споредуваат со соодветната РНА што следува (ако го поделиме множеството овие би биле тест

секвенците). Со ова правиме на некој начин "валидација" на можноста на техниката/моделот да предвидува

2. РЕФЕРЕНЦИ

*1.S*alama MA, Hassanien AE, Mostafa A. The prediction of virus mutation using neural networks and rough set techniques. EURASIP J Bioinform Syst Biol. 2016 May 13;2016(1):10. doi: 10.1186/s13637-016-0042-0. PMID: 27257410; PMCID: PMC4867776.

Дополнителни инфо

- Outbreak.info: https://outbreak.info/situation-reports/caveats
- Mutation explorer http://sars2.cvr.gla.ac.uk/cog-uk/
- Phylogeny of SARS-CoV-2: https://www.nature.com/articles/s41598-021-82938-2
- NCBI, SARS-CoV-2, complete genome https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2/
- NCBI, Spike protein https://www.ncbi.nlm.nih.gov/protein/YP_009724390.1/?report=fasta